Encyclopedia of

SURVEY Research Methods

EDITOR
Paul J. Lavrakas

1&2
VOLUME

Encyclopedia of
Survey Research Methods

Editorial Board

Encyclopedia of Survey Research Methods

EDITOR

Paul J. Lavrakas

Independent Consultant and Former Chief Research Methodologist for The Nielsen Company

VOLUME

1 & 2

SAGE

Los Angeles • London • New Delhi • Singapore • Washington DC

A SAGE Reference Publication

For information:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of survey research methods/editor, Paul J. Lavrakas.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4129-1808-4 (cloth)
1. Social surveys—Research—Encyclopedias. I. Lavrakas, Paul J.
HN29.E53 2008
300.72

Contents

With considerable gratitude to Seymour Sudman, David H. Weaver,
and Robert M. Groves for the key support they
provided at various times in my career.

List of Entries

Reader's Guide

The Reader's Guide is provided to assist readers in locating articles on related topics. It classifies articles into nine general topical categories: (1) Ethical Issues in Survey Research; (2) Measurement; (3) Nonresponse; (4) Operations; (5) Political and Election Polling; (6) Public Opinion; (7) Sampling, Coverage, and Weighting; (8) Survey Industry; and (9) Survey Statistics.

Respondent

Miscellaneous

Nonresponse

Item-Level

Outcome Codes and Rates

Unit-Level

Operations

General

Computerized-Response Audience Polling (CRAP)
Convention Bounce
Deliberative Poll
800 Poll
Election Night Projections
Election Polls
Exit Polls
Favorability Ratings
FRUGing
Horse Race Journalism
Leaning Voters
Likely Voter
Media Polls
Methods Box
National Council on Public Polls (NCPP)
National Election Pool (NEP)
National Election Studies (NES)
New York Times/CBS News Poll
900 Poll
Poll
Polling Review Board (PRB)
Pollster
Precision Journalism
Pre-Election Polls
Pre-Primary Polls
Prior Restraint
Probable Electorate
Pseudo-Polls
Push Polls
Rolling Averages
Sample Precinct
Self-Selected Listener Opinion Poll (SLOP)
Straw Polls
Subgroup Analysis
SUGing
Tracking Polls
Trend Analysis
Trial Heat Question
Undecided Voters

Public Opinion

Agenda Setting
Consumer Sentiment Index
Issue Definition (Framing)
Knowledge Gap
Mass Beliefs
Opinion Norms
Opinion Question
Opinions
Perception Question
Political Knowledge
Public Opinion
Public Opinion Research
Quality of Life Indicators
Question Wording as Discourse Indicators
Social Capital
Spiral of Silence
Third-Person Effect
Topic Saliency
Trust in Government

Sampling, Coverage, and Weighting

Adaptive Sampling
Add-a-Digit Sampling
Address-Based Sampling
Area Frame
Area Probability Sample
Capture–Recapture Sampling
Cell Phone Only Household
Cell Phone Sampling
Census
Clustering
Cluster Sample
Complex Sample Surveys
Convenience Sampling
Coverage
Coverage Error
Cross-Sectional Survey Design
Cutoff Sampling
Designated Respondent
Directory Sampling
Disproportionate Allocation to Strata
Dual-Frame Sampling
Duplication
Elements
Eligibility
Email Survey
EPSEM Sample
Equal Probability of Selection
Error of Nonobservation
Errors of Commission
Errors of Omission
Establishment Survey
External Validity
Field Survey
Finite Population
Frame
Geographic Screening

Hagan and Collier Selection Method
Half-Open Interval
Informant
Internet Pop-Up Polls
Internet Surveys
Interpenetrated Design
Inverse Sampling
Kish Selection Method
Last-Birthday Selection
List-Assisted Sampling
List Sampling
Log-in Polls
Longitudinal Studies
Mail Survey
Mall Intercept Survey
Mitofsky-Waksberg Sampling
Mixed-Mode
Multi-Mode Surveys
Multiple-Frame Sampling
Multiplicity Sampling
Multi-Stage Sample
n
N
Network Sampling
Neyman Allocation
Noncoverage
Nonprobability Sampling
Nonsampling Error
Optimal Allocation
Overcoverage
Panel
Panel Survey
Population
Population of Inference
Population of Interest
Post-Stratification
Primary Sampling Unit (PSU)
Probability of Selection
Probability Proportional to Size (PPS) Sampling
Probability Sample
Propensity Scores
Propensity-Weighted Web Survey
Proportional Allocation to Strata
Proxy Respondent
Purposive Sample
Quota Sampling
Random
Random-Digit Dialing (RDD)
Random Sampling
Ranked-Set Sampling (RSS)
Rare Populations
Registration-Based Sampling (RBS)
Repeated Cross-Sectional Design
Replacement
Representative Sample
Research Design
Respondent-Driven Sampling (RDS)
Reverse Directory Sampling
Rotating Panel Design
Sample
Sample Design
Sample Size
Sampling
Sampling Fraction
Sampling Frame
Sampling Interval
Sampling Pool
Sampling Without Replacement
Screening
Segments
Self-Selected Sample
Self-Selection Bias
Sequential Sampling
Simple Random Sample
Small Area Estimation
Snowball Sampling
Strata
Stratified Sampling
Superpopulation
Survey
Systematic Sampling
Target Population
Telephone Households
Telephone Surveys
Troldahl-Carter-Bryant Respondent Selection Method
Undercoverage
Unit
Unit Coverage
Unit of Observation
Universe
Wave
Web Survey
Weighting
Within-Unit Coverage
Within-Unit Coverage Error
Within-Unit Selection
Zero-Number Banks

Survey Industry

American Association for Public Opinion Research (AAPOR)

American Community Survey (ACS)
American Statistical Association Section on Survey Research Methods (ASA-SRMS)
Behavioral Risk Factor Surveillance System (BRFSS)
Bureau of Labor Statistics (BLS)
Cochran, W. G.
Council for Marketing and Opinion Research (CMOR)
Council of American Survey Research Organizations (CASRO)
Crossley, Archibald
Current Population Survey (CPS)
Gallup, George
Gallup Poll
General Social Survey (GSS)
Hansen, Morris
Institute for Social Research (ISR)
International Field Directors and Technologies Conference (IFD&TC)
International Journal of Public Opinion Research (IJPOR)
International Social Survey Programme (ISSP)
Joint Program in Survey Methods (JPSM)
Journal of Official Statistics (JOS)
Kish, Leslie
National Health and Nutrition Examination Survey (NHANES)
National Health Interview Survey (NHIS)
National Household Education Surveys (NHES) Program
National Opinion Research Center (NORC)
Pew Research Center
Public Opinion Quarterly (POQ)
Roper, Elmo
Roper Center for Public Opinion Research
Sheatsley, Paul
Statistics Canada
Survey Methodology
Survey Sponsor
Telemarketing
U.S. Bureau of the Census
World Association for Public Opinion Research (WAPOR)

Survey Statistics

Algorithm
Alpha, Significance Level of Test
Alternative Hypothesis
Analysis of Variance (ANOVA)
Attenuation
Auxiliary Variable
Balanced Repeated Replication (BRR)
Bias
Bootstrapping
Chi-Square
Composite Estimation
Confidence Interval
Confidence Level
Constant
Contingency Table
Control Group
Correlation
Covariance
Cronbach's Alpha
Cross-Sectional Data
Data Swapping
Design-Based Estimation
Design Effects (*deff*)
Ecological Fallacy
Effective Sample Size
Experimental Design
Factorial Design
Finite Population Correction (fpc) Factor
Frequency Distribution
F-Test
Hot-Deck Imputation
Imputation
Independent Variable
Inference
Interaction Effect
Internal Validity
Interval Estimate
Intracluster Homogeneity
Jackknife Variance Estimation
Level of Analysis
Main Effect
Marginals
Margin of Error (MOE)
Mean
Mean Square Error
Median
Metadata
Mode
Model-Based Estimation
Multiple Imputation
Noncausal Covariation
Null Hypothesis
Outliers
Panel Data Analysis
Parameter

Percentage Frequency Distribution
Percentile
Point Estimate
Population Parameter
Post-Survey Adjustments
Precision
Probability
p-Value
Raking
Random Assignment
Random Error
Raw Data
Recoded Variable
Regression Analysis
Relative Frequency
Replicate Methods for Variance Estimation
Research Hypothesis
Research Question
ρ (Rho)
Sampling Bias
Sampling Error
Sampling Variance
SAS
Seam Effect
Significance Level
Solomon Four-Group Design
Standard Error
Standard Error of the Mean
STATA
Statistic
Statistical Package for the Social Sciences (SPSS)
Statistical Power
SUDAAN
Systematic Error
Taylor Series Linearization
Test-Retest Reliability
Total Survey Error (TSE)
t-Test
Type I Error
Type II Error
Unbiased Statistic
Validity
Variable
Variance
Variance Estimation
WesVar
z-Score

About the General Editor

Paul J. Lavrakas, Ph.D., is a research psychologist (Loyola University of Chicago; 1975, 1977) and currently is serving as a methodological research consultant for several public-sector and private-sector organizations. He served as vice president and chief methodologist for Nielsen Media Research from 2000 to 2007. Previously, he was a professor of journalism and communication studies at Northwestern University (1978–1996) and at Ohio State University (1996–2000). During his academic career he was the founding faculty director of the Northwestern University Survey Lab (1982–1996) and the OSU Center for Survey Research (1996–2000). Prior to that he was a fifth-grade teacher in the inner-city of Chicago (1968–1972) and helped established a social science evaluation research unit for Westinghouse in 1976–1977. Among his publications, he has written two editions of a widely read book on telephone survey methodology (1987, 1993) and served as the lead editor for three books on election polling, the news media, and democracy (1991, 1995, 1999), as well as co-authoring four editions of *The Voter's Guide to Election Polls* (1996, 2000, 2004, 2008). He served as guest editor for a special issue of *Public Opinion Quarterly* on "Cell Phone Numbers and Telephone Surveys" published in December, 2007, and chaired a task force for the American Association for Public Opinion Research (AAPOR) which issued a report in 2008 on this topic (www.aapor.org). Dr. Lavrakas was a co-recipient of the 2003 AAPOR Innovators Award for his work on the standardization of survey response rate calculations, was named a Fellow of the Midwestern Association for Public Opinion Research (MAPOR) in 1995, and was recognized in 2007 with an Outstanding Career Achievement Award by the New York Association for Public Opinion Research (NYAAPOR). He has been elected twice to the AAPOR Executive Council as Program Chair (1997–1999) and Counselor at Large (2008–2010). Dr. Lavrakas was born in Cambridge, Massachusetts, and was educated in the public schools of Birmingham, Michigan. His undergraduate degree is from Michigan State University. His wife of 40 years, Barbara J. Lavrakas, and he live in Connecticut; their son, Nikolas J. Lavrakas, is a resident of Perth, Australia.

Contributors

Sowmya Anand
University of Illinois

Mindy Anderson-Knott
University of Nebraska–Lincoln

H. Öztas Ayhan
Middle East Technical University

Janice Ballou
Mathematica Policy Research

Badi H. Baltagi
Syracuse University

Laura Barberena
University of Texas at Austin

Kirsten Barrett
Mathematica Policy Research

Allen H. Barton
University of North Carolina at Chapel Hill

Danna Basson
Mathematica Policy Research

Michael P. Battaglia
Abt Associates, Inc.

Joseph E. Bauer
American Cancer Society

Christopher W. Bauman
Northwestern University

Sandra L. Bauman
Bauman Research

René Bautista
University of Nebraska–Lincoln

Patricia C. Becker
APB Associates, Inc.

Robert F. Belli
University of Nebraska–Lincoln

Mildred A. Bennett
The Nielsen Company

Pazit Ben-Nun
State University of New York, Stony Brook University

Sandra H. Berry
RAND

Marcus Berzofsky
RTI International

Jonathan Best
Princeton Survey Research International

Jelke Bethlehem
Statistics Netherlands

Matthew Beverlin
University of Kansas

David A. Binder
Statistics Canada

George F. Bishop
University of Cincinnati

Steven Blixt
Bank of America

Joel David Bloom
State University of New York, Albany

Stephen J. Blumberg
Centers for Disease Control and Prevention

Georgiy Bobashev
RTI International

Shelley Boulianne
University of Wisconsin–Madison

Ashley Bowers
University of Michigan

Diane Bowers
Council for American Survey Research Organization

Heather H. Boyd
University of Wisconsin–Extension

Luc Boyer
University of Waterloo

J. Michael Brick
Westat

Pat Dean Brick
Westat

Jonathan E. Brill
University of Medicine and Dentistry of New Jersey

Kimberly Diane Brown
The Nielsen Company

Trent D. Buskirk
Saint Louis University

Sarah Butler
National Economic Research Associates

Mario Callegaro
Knowledge Networks

Pamela Campanelli
The Survey Coach

Patrick J. Cantwell
U.S. Census Bureau

Xiaoxia Cao
University of Pennsylvania

Lisa Carley-Baxter
RTI International

Woody Carter
University of Chicago

Barbara L. Carvalho
Marist College

Rachel Ann Caspar
RTI International

Jamie Patrick Chandler
City University of New York

Haiying Chen
Wake Forest University

Young Ik Cho
University of Illinois at Chicago

Leah Melani Christian
Pew Research Center

James R. Chromy
RTI International

M. H. Clark
Southern Illinois University–Carbondale

Kathryn A. Cochran
University of Kansas

Jon Cohen
The Washington Post

Michael P. Cohen
Bureau of Transportation Statistics

Marjorie Connelly
The New York Times

Matthew Courser
Pacific Institute for Research and Evaluation

Brenda G. Cox
Battelle

Douglas B. Currivan
RTI International

Richard T. Curtin
University of Michigan

Gauri Sankar Datta
University of Georgia

Michael Edward Davern
University of Minnesota

Robert P. Daves
Daves & Associates Research

Bonnie D. Davis
Public Health Institute, Survey Research Group

Karen E. Davis
National Center for Health Statistics

Matthew DeBell
Stanford University

Femke De Keulenaer
Gallup Europe

Edith D. de Leeuw
Methodika

David DesRoches
Mathematica Policy Research, Inc.

Dennis Dew
Loyola University Chicago

Isaac Dialsingh
Pennsylvania State University

Lillian Diaz-Hoffmann
Westat

Bryce J. Dietrich
University of Kansas

Wil Dijkstra
Free University, Amsterdam

Don A. Dillman
Washington State University

Charles DiSogra
Knowledge Networks

Sylvia Dohrmann
Westat

Wolfgang Donsbach
Technische Universität Dresden

Katherine A. Draughon
Draughon Research, LLC

Arthur Lance Dryver
National Institute of Development Administration

Natalie E. Dupree
National Center for Health Statistics

Jennifer Dykema
University of Wisconsin

Asia A. Eaton
University of Chicago

Murray Edelman
Rutgers University

Mansour Fahimi
Marketing Systems Group

Moshe Feder
RTI International

Karl G. Feld
D3 Systems, Inc.

Howard Fienberg
CMOR

Agnieszka Flizik
BioVid

Amy Flowers
Market Decisions

E. Michael Foster
University of North Carolina at Chapel Hill

Kelly N. Foster
University of Georgia

Paul Freedman
University of Virginia

Marek Fuchs
University of Kassel

Siegfried Gabler
Universität Mannheim

Matthias Ganninger
Gesis-ZUMA

Cecilie Gaziano
Research Solutions, Inc.

Jane F. Gentleman
National Center for Health Statistics

Amy R. Gershkoff
Princeton University

Malay Ghosh
University of Florida

Ryan Gibb
University of Kansas

Homero Gil de Zuniga
University of Texas at Austin

Jason E. Gillikin
QSEC Consulting Group, LLC

Lisa M. Gilman
University of Delaware

Patrick Glaser
CMOR

Carroll J. Glynn
Ohio State University

John Goyder
University of Waterloo

Ingrid Graf
University of Illinois at Chicago

Eric A. Greenleaf
New York University

Thomas M. Guterbock
University of Virginia

Erinn M. Hade
Ohio State University

Sabine Häder
Gesis-ZUMA

John Hall
Mathematica Policy Research

Janet Harkness
University of Nebraska–Lincoln

Chase H. Harrison
Harvard University

Rachel Harter
University of Chicago

Douglas D. Heckathorn
Cornell University

Dirk Heerwegh
Katholieke Universiteit Leuven

Sean O. Hogan
RTI International

Allyson Holbrook
University of Illinois at Chicago

Gregory G. Holyk
University of Illinois at Chicago

Adriaan W. Hoogendoorn
Vrije Universiteit, Amsterdam

Lew Horner
Ohio State University

Joop Hox
Utrecht University

Michael Huge
Ohio State University

Larry Hugick
Princeton Survey Research International

Li-Ching Hung
Mississippi State University

Ronaldo Iachan
Macro International

Susan S. Jack
National Center for Health Statistics

Annette Jäckle
University of Essex

Matthew Jans
University of Michigan

Sharon E. Jarvis
University of Texas at Austin

Guillermina Jasso
New York University

E. Deborah Jay
Field Research Corporation

Timothy Johnson
University of Illinois at Chicago

David Ross Judkins
Westat

Karen Long Jusko
University of Michigan

Sema A. Kalaian
Eastern Michigan University

William D. Kalsbeek
University of North Carolina

Rafa M. Kasim
Kent State University

Randall Keesling
RTI International

Scott Keeter
Pew Research Center

Jenny Kelly
NORC at the University of Chicago

Courtney Kennedy
University of Michigan

John M. Kennedy
Indiana University

Timothy Kennel
U.S. Census Bureau

Kate Kenski
University of Arizona

SunWoong Kim
Dongguk University

Irene Klugkist
Utrecht University

Thomas R. Knapp
University of Rochester and Ohio State University

James R. Knaub, Jr.
U.S. Department of Energy

Gerald M. Kosicki
Ohio State University

Phillip S. Kott
USDA/NASS

John Kovar
Statistics Canada

Tom Krenzke
Westat

Frauke Kreuter
University of Maryland

Parvati Krishnamurty
NORC at the University of Chicago

Karol Krotki
RTI International

Dale W. Kulp
Marketing Systems Group

Richard Kwok
RTI International

Jennie W. Lai
The Nielsen Company

Dennis Lambries
University of South Carolina

Gary Langer
ABC News

Michael D. Larsen
Iowa State University

Paul J. Lavrakas
Independent Consultant and Former Chief Research Methodologist for The Nielsen Company

Geon Lee
University of Illinois at Chicago

Hyunshik Lee
Westat

Sunghee Lee
University of California, Los Angeles

Jason C. Legg
Iowa State University

Stanley Lemeshow
Ohio State University

Gerty Lensvelt-Mulders
Universiteit Utrecht

James M. Lepkowski
University of Michigan

Tim F. Liao
University of Illinois at Urbana-Champaign

Michael W. Link
The Nielsen Company

Jani S. Little
University of Colorado

Cong Liu
Hofstra University

Kamala London
University of Toledo

Geert Loosveldt
Katholieke Universiteit Leuven

Mary E. Losch
University of Northern Iowa

Thomas Lumley
University of Washington

Lars Lyberg
Statistics Sweden

Tina Mainieri
Survey Sciences Group, LLC

Aaron Keith Maitland
University of Maryland

Donald J. Malec
U.S. Census Bureau

Allan L. McCutcheon
University of Nebraska–Lincoln

Daniel G. McDonald
Ohio State University

John P. McIver
University of Colorado

Douglas M. McLeod
University of Wisconsin–Madison

Daniel M. Merkle
ABC News

Philip Meyer
University of North Carolina at Chapel Hill

Peter V. Miller
Northwestern University

Lee M. Miringoff
Marist College

Michael Mokrzycki
Associated Press

J. Quin Monson
Brigham Young University

Jill M. Montaquila
Westat

Christopher Z. Mooney
University of Illinois at Springfield

Geraldine M. Mooney
Mathematica Policy Research

Danna L. Moore
Washington State University

David W. Moore
University of New Hampshire

Jeffrey C. Moore
U.S. Census Bureau

Richard Morin
Pew Research Center

Patricia Moy
University of Washington

Mary H. Mulry
U.S. Census Bureau

Ralf Münnich
University of Trier

Joe Murphy
RTI International

Gad Nathan
Hebrew University of Jerusalem

Shannon C. Nelson
University of Illinois at Chicago

Thomas E. Nelson
Ohio State University

Traci Lynne Nelson
University of Pittsburgh

William L. Nicholls
U.S. Census Bureau (Retired)

Matthew C. Nisbet
American University

Andrew Noymer
University of California, Irvine

Barbara C. O'Hare
Arbitron, Inc.

Robert W. Oldendick
University of South Carolina

Randall Olsen
Ohio State University

Kristen Olson
University of Michigan

Diane O'Rourke
University of Illinois at Chicago

Larry Osborn
Abt Associates, Inc.

Ronald E. Ostman
Cornell University

Mary Outwater
University of Oklahoma

Linda Owens
University of Illinois at Urbana-Champaign

Michael Parkin
Oberlin College

Jennifer A. Parsons
University of Illinois at Chicago

Jeffrey M. Pearson
University of Michigan

Steven Pedlow
NORC at the University of Chicago

Chao-Ying Joanne Peng
Indiana University

Andy Peytchev
RTI International

Linda Piekarski
Survey Sampling International

Christine Guyer Pierce
The Nielsen Company

Kathy Pilhuj
Scarborough Research

Stephen R. Porter
Iowa State University

Frank Potter
Mathematica Policy Research

Kevin B. Raines
Corona Research, Inc.

Susanne Rässler
Otto-Friedrich-University Bamberg

Bryce B. Reeve
National Cancer Institute

Lance J. Rips
Northwestern University

José Elías Rodríguez
Universidad de Guanajuato

David James Roe
Survey Sciences Group

Jennifer M. Rothgeb
U.S. Census Bureau

Donald B. Rubin
Harvard University

Tamás Rudas
Eotvos Lorand University

Pedro Saavedra
ORC Macro

Adam Safir
RTI International

Joseph W. Sakshaug
University of Michigan

Charles T. Salmon
Michigan State University

Carla R. Scanlan
Independent Researcher

Fritz Scheuren
NORC at the University of Chicago

Michael F. Schober
New School for Social Research

Matthias Schonlau
RAND

Paul Schroeder
Abt SRBI

Tricia Seifert
University of Iowa

William R. Shadish
University of California at Merced

Dhavan V. Shah
University of Wisconsin–Madison

Jacob Shamir
Hebrew University of Jerusalem

Gary M. Shapiro
Westat

Joel K. Shapiro
Rockman et al.

Carol Sheets
Indiana University

Sarah Shelton
Saint Louis University

Charles D. Shuttles
The Nielsen Company

Samuel Shye
Hebrew University of Jerusalem

Richard Sigman
Westat

Carlos Nunes Silva
University of Lisbon

N. Clayton Silver
University of Nevada, Las Vegas

Jody Smarr
The Nielsen Company

Cary Stacy Smith
Mississippi State University

Tom W. Smith
NORC at the University of Chicago

Jolene D. Smyth
University of Nebraska–Lincoln

Elizabeth A. Stasny
Ohio State University

Jeffery A. Stec
CRA International

David Steel
University of Wollongong

Sonya K. Sterba
University of North Carolina at Chapel Hill

Kenneth W. Steve
Abt SRBI

John Stevenson
University of Wisconsin

James W. Stoutenborough
University of Kansas

John Tarnai
Washington State University

Charles Tien
City University of New York, Hunter College

Lois E. Timms-Ferrara
University of Connecticut

Trevor N. Tompson
Associated Press

Jeff Toor
San Diego State University

Roger Tourangeau
University of Maryland

Michael W. Traugott
University of Michigan

Alberto Trobia
University of Palermo

Norm Trussell
The Nielsen Company

Clyde Tucker
U.S. Bureau of Labor Statistics

Geoffrey R. Urland
Corona Research

Akhil K. Vaish
RTI International

Melissa A. Valerio
University of Michigan

Wendy Van de Kerckhove
Westat

Patrick Vargas
University of Illinois at Urbana-Champaign

Timothy Vercellotti
Rutgers University

Ana Villar
University of Nebraska–Lincoln

Penny Sue Visser
University of Chicago

Herbert F. Weisberg
Ohio State University

Eric White
University of Wisconsin

Rand R. Wilcox
University of Southern California

Rick L. Williams
RTI International

Gordon B. Willis
National Cancer Institute

Michael B. Witt
RTI International

Jonathan Wivagg
PTV DataSource

James Wolf
Indiana University at Indianapolis

Shapard Wolf
Arizona State University

Douglas A. Wolfe
Ohio State University

Daniel B. Wright
University of Sussex

Changbao Wu
University of Waterloo

Ting Yan
NORC at the University of Chicago

Y. Michael Yang
University of Chicago

Elaine L. Zanutto
National Analysts Worldwide

Elizabeth R. Zell
Centers for Disease Control and Prevention

Weiyu Zhang
University of Pennsylvania

Sonja Ziniel
University of Michigan

Mary B. Ziskin
Indiana University

Introduction

Survey research is a systematic set of methods used to gather information to generate knowledge and to help make decisions. By the second half of the 20th century, surveys were being used routinely by governments, businesses, academics, politicians, the news media, those in public health professions, and numerous other decision makers. It is not an exaggeration to state that accurate surveys have become a necessary condition for the efficient functioning of modern-day societies, and thus for our individual well-being.

Although there is a rich and expanding body of literature that has been produced mostly in the past half century about the myriad methods that are used by survey researchers, heretofore there has not been a compendium with information about each of those methods to which interested parties could turn, especially those new to the field of survey research. Thus, the purpose of the *Encyclopedia of Survey Research Methods* (*ESRM*) is to fill that gap by providing detailed (although not exhaustive) information about each of the many methods that survey methodologists and survey statisticians deploy in order to conduct reliable and valid surveys.

The Role of Methods and Statistics in the Field of Survey Research

A survey is often contrasted to a census, and the two use many of the same methods. However, whereas a census is intended to gather information about all members of a population of interest, a survey gathers information from only some of the population members, that is, from a sample of the population. Because a survey is more limited in how much information it gathers compared to a census with a comparable scope of variables needing to be measured, a survey is less costly than a census and often is more accurate and timelier. Due to its smaller scope, it is easy to understand why a survey is less costly and timelier than a census, but it may surprise some to learn that a survey can be more accurate than a census. That is the case because a census often is a daunting enterprise that cannot be conducted accurately across an entire population. At far less cost than a census, a survey can sample a representative subset of the population, gain a very high response rate, gather data on the same variables a census measures, and do so much more quickly than a census. Thus, given the finite resources available for information gathering, survey researchers often can allocate those resources much more effectively and achieve more accurate results than those conducting a census on the same topic.

There are two primary defining characteristics of a survey. One is that a sample is taken from the population and the other is that a systematic instrument—most often a structured questionnaire—is used to gather data from each sampled member of, or unit in, the population.

However, the general methods of "surveying" are used in many ways other than their well-recognized manifestations in survey research. At the broadest level, humans are always "sampling" the physical and social environments in which they live, "gathering" information in mostly unstructured ways, and "analyzing" the information to reach decisions, albeit often imperfectly. And although survey research is considered a quantitative approach for gathering information, "surveying" is routinely performed by qualitative researchers, even if many may not think of themselves as using survey methods. That is, qualitative research "samples" some members from a population of interest so as to gather information from or about them. This includes qualitative research that uses content analysis, focus groups, observational methods, ethnographic methods, and other quasi-scientific information-gathering approaches.

Whether the samples drawn for qualitative research are representative, and whether the information-gathering means are reliable, is not the primary issue here. Instead, the issue is that qualitative research relies on "survey methods" even if many who practice it have had no rigorous training in those methods. Also, there are many fields of inquiry in the behavioral sciences that utilize survey methods even if they do not recognize or acknowledge that is what is being done. For example, many psychologists draw samples and use questionnaires to gather data for their studies, even if they do not think of themselves as survey researchers or have not had rigorous training in survey methods. The same holds for many political scientists, economists, sociologists, criminologists, and other social scientists, as well as many public health researchers.

Accuracy Versus Error in Survey Research

The goal of a good survey is to utilize available resources so as to gather the most accurate information possible. No survey researcher should (or can) claim that a survey is entirely without error, that is, that it is perfectly accurate or valid. Instead, what survey researchers realistically can strive for is to gather as accurate information as possible with available resources—information that has the smallest amount of "total survey error." Ideally this will result in an amount of error that is "negligible," that is, ignorable, for the decision-making purposes that the survey is to serve. For example, the senior executives of a corporation do not need to know exactly what proportion of the population is likely to purchase their new product. Rather, they can make a confident decision about whether to proceed with introducing the product on the basis of survey estimates that are accurate within a tolerable (negligible) level of "error."

Broadly speaking, error in surveys takes two forms: variance and bias. Variance refers to all sources of imprecision that may affect survey data. Variance is a random form of error, which can be likened to "noise," and there are many approaches that can be used to reduce its size or to measure its size. Bias is a constant form of error and thus is directional: positive or negative. In some cases, bias leads to survey data that underestimate what is being measured, whereas in other cases, bias leads to overestimates. On occasion, different types of biases cancel out their own separate effects on survey estimates, but often it is very difficult for researchers to know when this has occurred. There are many methods that researchers can use to try to avoid bias, as well as many that can estimate the presence, size, and nature of bias. But all of these methods add costs to survey projects, and in many cases these added costs are great indeed.

In designing a survey, researchers should strive to allocate available resources so as to reduce the impact of likely errors, measure the size of the errors, or both, and then take that knowledge into account when drawing conclusions with the data generated by the survey. To accomplish this, researchers must be well aware of the various survey methods that can be used, and then they must select the ones that are most likely to achieve the most beneficial balance of both these goals. This requires survey researchers to constantly make trade-offs in choosing the "best" methods for their particular survey project. Allocating too many resources for one type of method will limit what can be allocated for other methods. If the first method addresses a source of error that is smaller in size than what will result from another source of error, then the allocation choice will have proven counterproductive in addressing total survey error concerns.

There are numerous types of possible errors that can occur with any survey, and it is the purpose of survey methods to address, and ideally avoid, all of these errors. It has been found useful to categorize these possible errors into a limited number of "types," which logically follow the chronology of planning, conducting, and analyzing a survey. The following sequence of questions summarizes this typology:

1. What is the population that must be studied, and how well will this population be "covered" (represented) by the frame (i.e., list) from which the sample will be drawn? This concerns coverage error.
2. How large will be the sample of frame members chosen for measurement, and what sampling design will be deployed to select these members? This concerns sampling error.
3. Among all the sampled members of the population, how will a high response rate be achieved, and will the nonresponders differ from responders in non-negligible ways on the variables of interest? This concerns nonresponse error.
4. What variables will be measured, and by what means will accurate data be gathered from the responding sample? This concerns specification

error, question-related measurement error, interviewer-related measurement error, respondent-related measurement error, and mode-related measurement error.

5. How will the data be processed, weighted, and analyzed? This concerns adjustment error and processing error.

Rationale for the *Encyclopedia of Survey Research Methods*

There is a considerable amount of existing literature on survey research and the methods that are used to conduct surveys. This exists in book form, in handbook chapters, in journal articles, in published conference proceedings, as well as an expanding body of otherwise unpublished works available via the Internet. The field is growing rapidly, both in the scope of what is known about survey methods and the importance this knowledge plays. However, to date, there has not existed a compendium to which interested parties, especially those without advanced knowledge of survey methods, can turn to learn about the great many topics that comprise the field of survey methodology.

The purpose of the *ESRM* is to fill that gap by being comprehensive in its coverage of the field, although not exhaustive in its explanation of any one topic. By providing more than 600 entries about important topics across the entirety of survey methodology, the encyclopedia serves as a "first place" to turn for those who need to learn about an aspect of survey methodology. The text of the entries in the encyclopedia will provide all the information that many users will need and desire. However, for those who want more information about a particular topic, the cross-referencing associated with nearly all of the entries provides these readers with guidance on where else to turn in the encyclopedia for additional information. And, for those who need still more information on a topic, essentially every entry provides a road map to additional readings.

Content and Organization of the Encyclopedia

The *ESRM* provides information about nearly all types of survey methods and survey errors. The more than 600 entries in the encyclopedia fall out across the following categories, which are listed in full detail in the Reader's Guide:

Ethics. These entries address a wide range of ethical matters that affect survey research, such as confidentiality, anonymity, debriefing, informed consent, voluntary participation, disclosure, and deception. Although addressing ethical issues complicates the methods that survey researchers must use and adds to the costs of surveys, it is critical that the survey research profession earn and maintain credibility and respect through observing strong ethical principles.

Measurement. The measurement entries focus on all nonoperational aspects of data collection, from conceptualization of the questionnaire through data collection and the effects that respondents have on data quality. This includes a wide range of entries covering question-related topics (such as closed-ended question, double-negatives, graphical language, mutually exclusive, question stem, and self-reported measure), interviewer-related topics (such as conversational interviewing, interviewer neutrality, nondirective probing, and standardized survey interviewing), respondent-related topics (such as acquiescence response bias, comprehension, telescoping, nondifferentiation, primacy effect, and satisficing), and mode-related topics.

Nonresponse. The entries on the topic of nonresponse are among the most important in the encyclopedia, as many scholars and practitioners regard nonresponse as the most daunting challenge facing survey research. This set of entries includes ones related to unit nonresponse, item nonresponse, and response outcomes and rates. These entries include incentives, leverage-saliency theory, completion rate, differential attrition, nonignorable nonresponse, missing data, refusal conversion, and tailoring.

Operations. These entries focus on a wide range of operational and technical topics related to the various modes of data collection, but predominantly surveys that are conducted in person (such as computer-assisted personal interviewing, control sheet, field work, and residence rules) and via the telephone (such as answering machine messages, calling rules, Federal Trade Commission (FTC) regulations, number portability, and predictive dialing). This grouping also includes operational entries related to surveys that gather data via interviewers (such as interviewer training, interviewer monitoring, and interviewer debriefing)

Political and Election Polling. This group includes survey methods that are specific to election-related and other types of political polling. These entries include measurement topics (such as approval ratings, convention bounce, leaning voters, and probable electorate), media-related topics (such as election night projections, horse race journalism, and precision journalism) and types of election or political surveys (such as deliberative polls, exit polls, pre-primary polls, and tracking polls).

Public Opinion. The entries in the public opinion grouping focus on a wide range of theoretical matters that affect the understanding of public opinion, with special attention to the methodological issues that are related to each theoretical concept. This set of entries includes agenda setting, knowledge gap, spiral of silence, third-person effect, and trust in government.

Sampling, Coverage, and Weighting. This group covers a large and broad set of entries, many of which are interrelated to sampling, coverage, and weighting, such as address-based sampling, cell phone sampling, coverage error, designated respondent, finite population, interpenetrated design, Neyman allocation, poststratification, quota sampling, replacement, sample size, undercoverage, and zero-number banks.

Survey Industry. The entries in the survey industry grouping include ones describing major survey professional organizations (such as AAPOR, CMOR, and CASRO), major academic-based survey organizations and government-based survey agencies (such as NORC, ISR, Bureau of Labor Statistics, and Statistics Canada), major figures in the history of survey research (such as Elmo Roper, Leslie Kish, Morris Hansen, and George Gallup), major U.S. government surveys (such as the Behavioral Risk Factor Surveillance System, the Current Population Survey, and the National Health Interview Survey), and major survey research periodicals (such as *Public Opinion Quarterly,* the *Journal of Official Statistics,* and the *International Journal of Public Opinion Research*).

Survey Statistics. The survey statistics grouping covers a diverse spectrum of statistical concepts and procedures that survey researchers use to help analyze and interpret the data that surveys generate. These include balanced repeated replication, control group, design-based estimation, hot-deck imputation, margin of error, outliers, perturbation methods, random assignment, sampling variance, test–retest reliability, and Type I error.

Despite the efforts of the editor, the members of the Editorial Board, and the many contributors who suggested new topics for inclusion, not every topic that someone interested in survey methods may seek knowledge about is included in this first edition of the *ESRM.* An encyclopedia such as this is bound to disappoint some who rightly believe that an important topic is missing. The editor and publisher can only hope that no key topic in the field is missing and that few other truly important topics are missing. When there is an opportunity for a second edition, those gaps can be corrected.

Readers will also find some degree of overlap in some of the topic areas. This is believed to be preferable because readers generally will be better helped by encountering too much information on a topic than too little. Similarly, some related topics have been written by contributors who are not fully in agreement with each other about the broader topic area. This too is viewed to be beneficial to readers, as it demonstrates where uncertainties and ambiguities in the field exist in the understanding and the valuing of a specific survey method.

How the Encyclopedia Was Created

A remarkably large number of people made this work possible by contributing to it in many different ways. This includes the editor, our Editorial Board members, editorial and administrative staff at both Sage Publications and The Nielsen Company, and the more than 320 individuals throughout the world who contributed the more than 640 entries that appear in these two volumes.

Due in part to my nearly 30 years of experience as a survey researcher, both as an academic and in the private sector, I was approached by Sage in late 2004 and invited to serve as editor of the encyclopedia. At that time I was employed as chief research methodologist for The Nielsen Company. Sage also asked if Nielsen might serve as "corporate sponsor" for the encyclopedia. I approached Nielsen's chief research officer and readily secured his support for my involvement and the company's endorsement of the venture.

Work on the encyclopedia followed a logical process, whereby (a) the list of entries was assembled; (b) contributors for each entry were identified; (c) individual entries were submitted to the Web-based Sage Reference Tracking (SRT) system; (d) draft contributions were reviewed, edited, and revised as needed; and (e) revised entries were finalized by members of the Editorial Board and me. Sage editors performed additional editing, passed the text along to Sage's production departments, and then I did the final review of the page proofs. Mistakes that remain are mine, and with such a daunting project to manage, there are bound to be at least a few. For these I apologize to the affected contributors and readers.

To build the list of entries, I started by reviewing a comprehensive glossary of methodological survey terms that was assembled for one of my previous publications. Some of these topics were kept and others dropped. Using my own knowledge and experience, I added to this draft list and found that I had approximately 400 topics. I grouped the entries on the list into the categories that were used to organize the Reader's Guide (see groupings described previously). For each of these categories I had chosen Editorial Board members with expertise in that subject area. I circulated the draft list of entries in each category to the Editorial Board member(s) assigned to that category and asked for their input of additional entry titles. This process raised the number of entries on the list to approximately 550. The Editorial Board members and I identified contributors to invite for the majority of these entries. Using Sage's versatile and comprehensive SRT system, email invitations were sent. The vast majority of first invitations were accepted. In some cases, coauthors were proposed by the first author. In many cases where the original invitee could not accept, he or she recommended someone else with expertise in the topic area and that person was invited.

For those entries for which I was unsure whom to invite, I posted a series of emails onto two listserves, inviting qualified contributors to volunteer for the unassigned entries: the American Association for Public Opinion Research listserve, AAPORnet, and the Survey Research Methods Section of the American Statistical Association listserve, SRMSnet. These postings were disseminated further by users of those listserves to their colleagues and to other listserves. This approach, which originally I had not anticipated doing, turned out to be a windfall for the *ESRM*, as it brought out a wide array of international experts in survey research who would not otherwise have had an opportunity to contribute due to my own limitations in heretofore not knowing them well or at all. I cannot thank enough the members of AAPOR and SRMS-ASA, as well the contributors not affiliated with either organization, for their generous efforts to benefit the *ESRM*.

A final source of additional entry titles came from contributors themselves. As they were writing their entries and reviewing the list of entries on the SRT, they would contact me with recommendations for new entries to be added. As these recommendations came in, the Editorial Board and I made a case-by-case decision about whether the suggestion fit the scope of the *ESRM*, and in most cases it did.

Acknowledgments

I would like to begin by thanking Sage Publications for believing that there should be an *Encyclopedia of Survey Research Methods* and that I was a good choice to serve as its editor. Here Lisa Cuevas Shaw, acquisitions editor at Sage, played a major role. I am indebted to Diana Axelsen, the developmental editor at Sage with whom I worked most closely during the final 2 years of the project, for her intelligence, guidance, encouragement, patience, and friendship. I also thank Letty Gutierrez, reference systems manager at Sage, for the numerous occasions that she "fixed" things in the SRT that I was not able to. At the copyediting and production stages, I am especially grateful to the conscientiousness, editing abilities, commitment, and flexibility of Tracy Buyan (production editor), Colleen Brennan (copy editor), and Pam Suwinsky (copy editor). There were many others at Sage who worked hard and intelligently to make this encyclopedia possible, but I am especially thankful to those who created, maintained, and updated the SRT, which provided the Web-based platform that managed almost all the invitations, submissions, reviews, and revisions.

I also am indebted to Jody Smarr, the administrative staff member at The Nielsen Company, who was assigned to work with me during the last 2 years of the project, including the last 13 months after I ended my employment with the company. Ms. Smarr's intelligence, organization, reliability, and calm demeanor will always be remembered and appreciated. I also thank Paul Donato, chief research officer at Nielsen, for committing that the company would be supportive

of the venture and for following through on that commitment without hesitation. As the largest and most-profitable survey research organization in the world, it is highly fitting that Nielsen has served as the "corporate sponsor" of the *ESRM*.

Each and every member of the Editorial Board was central to the success of the project, and I appreciate all that each of them did. They made suggestions of topics to be added to the entry list; they recommended contributors and, many times, encouraged these persons to accept their invitations; they reviewed entries; and they also wrote entries themselves. Michael Link, originally of the Centers for Disease Control and later of Nielsen, helped with the entries in the categories of Ethics and Operations. Linda Piekarski of Survey Sampling International helped with Operations. Edith de Leeuw of Methodika helped with Nonresponse. Dan Merkle of ABC News and Mike Traugott of University of Michigan helped with Election and Political Polling. Carroll Glynn of Ohio State University helped with Public Opinion. Mike Battaglia of Abt Associates, Inc., Trent Buskirk of St. Louis University, Elizabeth Stasny of Ohio State University, and Jeff Stec of CRA International helped with Sampling, Coverage, and Weighting and Survey Statistics. Allyson Holbrook of University of Illinois at Chicago and Peter Miller of Northwestern University helped with Measurement.

Sage and I are also indebted to each of the contributors. Without their expertise, commitment, and belief that the *ESRM* would be a valuable addition to the field of survey research, the project could not have come to fruition. Survey methodologists and survey statisticians are a generous lot. They routinely give of their own time to help the field. They share knowledge freely for the sake of science. They want to make the world a better place, in part through their abilities and interests to use surveys to generate reliable and valid knowledge. There is one researcher, J. Neil Russell, who exemplifies this ethos, who could not be listed formally as a contributor in the *ESRM* because of employment-related reasons but who nevertheless was a coauthor for some of the entries. It is this level of commitment to the field of survey research that all *ESRM* contributors and I are proud to strive for.

Paul J. Lavrakas, Ph.D.
Stamford, Connecticut

A

ABC NEWS/*WASHINGTON POST* POLL

ABC News and *The Washington Post* initiated their polling partnership on February 19, 1981, announcing an 18-month agreement to jointly produce news surveys on current issues and trends. More than 25 years, 475 surveys, and 500,000 individual interviews later, the partnership has proved an enduring one. Their first shared survey—known as the ABC/*Post* poll to viewers of ABC News, and the *Post*/ABC survey to readers of the *Post*—focused on newly elected President Ronald Reagan's tax- and budget-cutting plans. While their work over the years has covered attitudes on a broad range of social issues, ABC and the *Post* have focused their joint polling primarily on politics and elections.

The two organizations consult to develop survey subjects, oversee methodology and research, and write questionnaires; each independently analyzes and reports the resulting data. Sampling, field work, and tabulation for nearly all ABC/*Post* polls have been managed from the start by the former Chilton Research Services, subsequently acquired by the multi-national research firm Taylor Nelson Sofres.

In addition to full-length, multi-night surveys, ABC and the *Post* have shared other polls designed to meet news demands, including one-night surveys (e.g., immediately after the terrorist attacks of September 11, 2001); daily pre-election tracking polls, in which the *Post* joined ABC as of 2000; and a weekly consumer confidence survey, in which the *Post* in 2005 joined an ABC effort ongoing since 1985.

The *Post* has been polling on its own since 1975, ABC since 1979. Their partnership was created by Dick Wald, senior vice president of ABC News, and his friend Ben Bradlee, the *Post*'s editor. Wald pitched the idea at lunch. Bradlee said, "Okay. You have a deal," he recalled. "We just shook hands. There was no contract, no paper, no anything else."

Jeffrey Alderman was longtime director of the survey for ABC, replaced in 1998 by Gary Langer. Barry Sussman directed for the *Post*, replaced in 1987 by Richard Morin, who in turn was succeeded in 2006 by Jonathan Cohen, then ABC's assistant polling director.

The news organizations also conduct polls on their own and with other partners. In 2005, ABC won the first news Emmy Award to cite a public opinion poll, for its second national survey in Iraq, on which it partnered with the BBC, the German network ARD, and *USA Today*. ABC also won the 2006 Iowa/Gallup award and 2006 National Council on Public Polls award for its polling in Iraq and Afghanistan; the *Post* won the 2007 Iowa/Gallup award for its survey focusing on black men in America, a poll it conducted with the Henry J. Kaiser Family Foundation and Harvard University.

Their joint polling nonetheless has been the most consistent feature of both organizations' efforts to cover the beat of public opinion. A search of the Factiva news archive for the 20 years through mid-2007 found 11,266 media references to ABC/*Post* polls, far surpassing references to any of the other ongoing news-sponsored public opinion surveys.

Gary Langer

See also Media Polls; *New York Times*/CBS News Poll

ACCESS LINES

An access line is a telecommunications link or telephone line connecting the central office or local switching center of a telephone company to the end user. Access lines are sometimes referred to as local routing numbers (LRNs), wireline loops, or switched access lines, and they do not include telephone numbers used for wireless services. Access lines provide access to a residence or business over twisted-pair copper wire, coaxial cable, or optical fiber. The Federal Communications Commission reported that as of December 31, 2005, there were approximately 175.5 million switched access lines in the United States. Access lines are normally assigned in prefixes or 1000-blocks classified by Telcordia as POTS ("Plain Old Telephone Service"), and most frames used for generating telephone samples are restricted to POTS prefixes and 1000-blocks.

Approximately two thirds of all access lines connect to a residence, which suggests that two thirds of working numbers in a telephone sample should be residential. Many business access lines are in dedicated prefixes or banks and do not appear in a list-assisted random-digit dialing (RDD) telephone sample. However, since a single business will frequently have multiple access lines, such as rollover lines, direct inward dial lines, fax lines, and modem lines, those access lines that are not in dedicated banks will appear in an RDD sample, substantially increasing the number of ineligible units.

A household also may have more than one access line. Over the years some households added additional access lines for children or home businesses. The increased use of home computers and residential fax machines in the 1990s further increased the number of residences with two or more access lines. Because multiple lines meant multiple probabilities of selection for a household, telephone surveys have regularly included a series of questions designed to determine the number of access lines or telephone numbers in a household. Between 1988 and 2001, the percentage of households with one or more nonprimary lines grew from approximately 2% to 26%. Dedicated computer lines have caused problems for telephone survey researchers, since these lines typically ring but are never answered, resulting in unknown eligibility status. Consequently, survey questions designed to determine the number of access lines have had to be adjusted to determine the number of lines that would ever be answered. Since 2001, the number of residential access lines has been declining. Many households have given up second lines and moved from dial-up Internet service to broadband service. Other households have opted to substitute wireless service for wireline service for some or all of their access lines. Current estimates suggest that, in 2007, 13% of households had only wireless telephone service.

Although an access line usually connects to a business or a residence, it may also connect to a pay phone, fax machine, or modem. Access lines can be used to obtain directory assistance, connect to Internet service providers, and order special programming from a cable or satellite service provider. An access line may not always connect to a specific location or device. Call forwarding allows a telephone call to be redirected to a mobile telephone or other telephone number where the desired called party is located. An access line can also be ported to another access line. Local number portability is the ability of subscribers to keep their existing telephone numbers when changing from one service provider to another. Porting requires two 10-digit numbers or access lines for each telephone number that is switched. One is the original subscriber number and the other is the number associated with the switch belonging to the new carrier. Finally, nascent Voice over Internet Protocol (VoIP) technologies and "virtual" phone numbers allow an access line to connect to either a telephone or computer that may or may not be located at the physical address associated with that access line or switch.

Linda Piekarski

See also Call Forwarding; Cell Phone Only Household; Eligibility; Federal Communications Commission (FCC) Regulations; Hit Rate; Number Portability; Prefix

ACQUIESCENCE RESPONSE BIAS

Acquiescence response bias is the tendency for survey respondents to agree with statements regardless of their content.

Acquiescence response bias could influence any question in which the response options involve confirming a statement, but it may be particularly problematic with agree–disagree questions. Although many guides on writing survey questions recommend avoiding agree–disagree questions, such questions are ubiquitous in survey instruments. An agree–disagree question asks respondents to report whether they agree or disagree with a statement. For example, respondents might be asked whether they agree or disagree with the statement, *It is important for the president to be a person of high moral character.* Acquiescence response bias is problematic because the interpretation of an "agree" response is very different if respondents are asked whether they agree or disagree with the posited statement than if they are asked whether they agree or disagree with the statement, "It is not important for the president to be a person of high moral character."

There are a number of explanations for acquiescence response bias. One explanation is that acquiescence response bias occurs partly due to social norms to be polite. Consistent with this, acquiescence response bias is stronger among cultures that put a high value on politeness and deference. Satisficing theory also provides an account for acquiescence response bias. Satisficing theory suggests that although survey researchers hope respondents will answer questions carefully and thoughtfully, respondents may not always be able or motivated to do so. Instead, they may shift their response strategies to minimize effort while providing a satisfactory response to the survey question (known as *satisficing*). One such strategy involves agreeing with assertions made by the interviewer. Satisficing theory also posits that satisficing is more likely when respondents' ability and motivation is low and when question difficulty is high. Thus, acquiescence response bias is likely to be strongest among respondents low in ability and motivation and for questions that are more difficult, a perspective that is supported by research studying acquiescence response bias.

There are also a number of strategies researchers use to avoid or control for acquiescence response bias. One such strategy is to include multiple items to measure a construct of interest, approximately half of which are worded so that the "agree" response indicates one position and the other half worded so that the "agree" response indicates the opposite position. For example, respondents might be asked whether they agree or disagree with the statement, "It is important for the president to be a person of high moral character," and then later asked whether they agree or disagree with the statement, "It is not important for the president to be a person of high moral character." If respondents exhibit acquiescence response bias and agree with both statements, their answers to these two questions cancel each other out.

There are at least three problems with this approach. First, it requires that survey researchers use a large number of redundant questions. This strategy is inefficient and it may be frustrating to respondents. Second, if researchers average responses across questions, this strategy results in "acquiescers" being given scores in the middle of the dimension, and it is not clear that this is appropriate or valid. Finally, as in the case discussed earlier, it sometimes results in respondents being asked whether they agree or disagree with a negative statement (e.g., "It is *not* important . . ."). This may be confusing to respondents, as disagreeing with this statement involves a double negative (respondents are reporting that they disagree that it is not important). This is a particular concern because not all languages treat double negatives in the same way, and agree–disagree questions about negative statements may therefore be particularly confusing for respondents for whom English is not their primary language or if questions are translated into other languages.

Another strategy for dealing with acquiescence response bias in agree–disagree questions involves rewriting all questions so that each question requires respondents to report directly about the dimension of interest. For example, the previous series of questions about the importance of the president's moral character could be rewritten to read, "How important do you believe it is for the president to have a strong moral character: extremely important, very important, somewhat important, a little important, or not at all important?" This strategy also allows researchers to follow experts' recommendations to avoid agree–disagree questions.

Allyson Holbrook

See also Likert Scale; Response Bias; Satisficing

Further Readings

Javeline, D. (1999). Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*, *63*, 1–28.
Johnson, T. P., Kulesa, P., Cho, Y. I. , & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264–277.
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.
Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58–88.
Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.
van de Vijver, F. J. R. (2004, October). *Toward an understanding of cross-cultural differences in acquiescence and extremity scoring.* Paper presented at the Sheth Foundation/Sudman Symposium on Cross-Cultural Survey Research, University of Illinois at Urbana-Champaign.

Adaptive Sampling

Adaptive sampling is a sampling technique that is implemented while a survey is being fielded—that is, the sampling design is modified in real time as data collection continues—based on what has been learned from previous sampling that has been completed. Its purpose is to improve the selection of elements during the remainder of the sampling, thereby improving the representativeness of the data that the entire sample yields.

Background

The purpose of sampling is to learn about one or more characteristics of a population of interest by investigating a subset, which is referred to as a *sample*, of that population. Typical population quantities of interest include the population mean, total, and proportion. For example, a population quantity of interest might be the total number of people living in New York City, their average income, and so on. From the sample collected, estimates of the population quantities of interest are obtained. The manner in which the sample is taken is called a *sampling design*, and for a sampling design various estimators exist. There is a multitude of sampling designs and associated estimators. Many factors may be considered in determining the sampling design and estimator used. The main objective is to use a sampling design and estimator that yield the most precise and accurate estimates utilizing the resources available. In conventional sampling designs and estimators, the sample is taken without regard to the unit values observed. That is, the observations obtained during sampling are not used in any manner to alter or improve upon future sample selections.

In adaptive sampling, on the other hand, the sampling selections depend on the observations obtained during the survey. In this sense, adaptive sampling designs are adaptive in that, while sampling, the remaining units to be sampled may change according to previously observed units. For design-based sampling, adaptive sampling could be a more efficient design to improve the inference and also increase the sampling yield. For model-based sampling, it has been shown that the optimal sampling strategy should be an adaptive one in general under a given population model.

Adaptive sampling designs have been used in various disciplines, including the ecological, epidemiological, environmental, geographical, and social sciences.

Adaptive Cluster Sampling

Adaptive cluster sampling (ACS) is a subclass of adaptive sampling designs. There has been considerable research within the field of adaptive sampling, utilizing ACS designs and their associated estimators. There are variations of ACS, such as stratified ACS, systematic ACS, ACS without replacement of clusters, and so on. The ACS designs are often more efficient than their conventional counterparts on clustered, or patched, populations. Typically this type of sampling design—ACS—is not only more efficient but also more useful for obtaining observations of interest for rare, hard-to-find, or elusive clustered populations. For example, there are various species of animals known to travel in groups and that are rare, such as whales. Through ACS, more whales may be captured in the sample than through conventional sampling techniques using a comparable final sample size of geographic locations. For surveys focused on elusive or hidden populations, such as individuals who are intravenous drug users, or HIV-positive individuals, ACS can aid greatly in increasing the number of individuals in the survey who meet the desired characteristics.

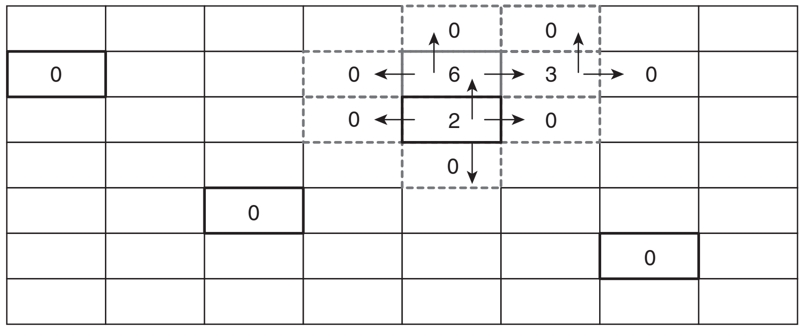

Figure 1 A final sample using ACS design with an initial simple random sample without replacement of size $n = 4$ from a population of size $N = 56$

Before a sampling commences, the condition to adaptively add units into the sample must be defined. Then an initial sample is drawn by some conventional sampling design. For example, for the original ACS, an initial sample is selected by simple random sampling with or without replacement. For stratified ACS, an initial sample is selected by stratified sampling; and for systematic ACS, an initial sample is selected by systematic sampling. With ACS, after the initial sample has been selected, units "in the neighborhood" of units in the sample that meet the predefined condition are added to the sample. If any of the adaptively added units meet the desired condition, then units in their neighborhood are added, and this process continues until no adaptively added units meet the predefined condition.

A neighborhood must be defined such that if unit i is in the neighborhood of unit j, then j is in the neighborhood of unit i. In addition to this restriction, a neighborhood can be defined in many ways, such as by spatial proximity, social relationship, and so on. All units within the neighborhood of one another that meet the predefined condition are called a *network*. Units that are in the neighborhood of units meeting the predefined condition but do not meet the predefined condition are called *edge units*. A network plus its associated edge units are called a *cluster;* thus the name *adaptive cluster sampling*. Only after the entire cluster has been observed is the size of a network containing units meeting the condition known. Often researchers do not desire to sample edge units, as they do not meet the predefined condition; unfortunately, which unit will be on the "edge" of a network remains unknown until after the unit has been observed. In addition, units not meeting the condition, including edge units, are networks of size 1. Figure 1 is an example of a final sample from an ACS, with an initial simple random sample without replacement taken from a forest partitioned into $N = 56$. The objective is to estimate the number of wolves in the forest. The condition to adaptively add neighboring units is finding one or more wolves in the unit sampled. The neighborhood is spatial and defined as north, south, east, and west. The initial sample is of size $n = 4$, represented by the dark bordered units. The units with a dotted border are adaptive added units. The adjacent units with the values 2, 6, 3 form a network of size 3. The units with a dotted border and a value of zero are edge units. The edge units plus the latter network of size 3 form a cluster. The edge units and the other units in the sample with a value of zero are networks of size 1.

In ACS, networks are selected with unequal probability. In typical unequal probability sampling, the probability of units included in the sample is determined before sampling begins. The typical estimators in ACS can be viewed as a weighted sum of networks, where the size of the network and whether the network was intersected in the initial sample is used to calculate the weights. Networks that are also edge units can enter into the final sample by being intersected in the initial sample or by being adaptively added, whereas other networks must be intersected in the initial sample. For the latter reason, the typical estimators do not incorporate edge units not intersected in the initial sample. Some estimators have been derived using the Rao-Blackwell theorem, which can incorporate edge units in the final sample but not in the initial sample.

For various reasons, when taking an ACS, it is often not feasible to sample the entire cluster; for example, because there are too many units to sample, cost-related issues, nonresponse, and so on. For this reason there has been research on estimation of the population quantities of interest in ACS when the entire cluster cannot be sampled, such as a restricted ACS design. A restricted ACS design is similar to a typical ACS design except that sampling stops after a predetermined number of units have been observed in the sample, regardless whether or not an entire network has been sampled. Biased and unbiased estimators have been derived for a restricted ACS design.

Adaptive Web Sampling

Recent research within adaptive sampling is the development of a new class of adaptive sampling designs called adaptive web sampling (AWS). The class of AWS designs is useful for sampling in network and spatial settings. A major distinction between ACS and AWS is that in ACS, units in the neighborhood of a sampled unit meeting a predefined condition are to be automatically adaptively added, whereas in AWS this is not so. In AWS it is possible to assign a probability to adding units adaptively in the neighborhood of units meeting a predefined condition. In the latter sense, AWS may be viewed as more flexible than ACS.

Arthur Lance Dryver

See also Design-Based Estimation; Model-Based Estimation; Probability of Selection; Sample; Sample Design; Sampling Without Replacement

Further Readings

Chao, C.-T., & Thompson, S. K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics*, *12*, 517–538.

Dryver, A. L., & Thompson, S. K. (2006). Adaptive sampling without replacement of clusters. *Statistical Methodology*, *3*, 35–43.

Selahi, M. M., & Seber, G. A. F. (2002). Unbiased estimators for restricted adaptive cluster sampling. *Australian and New Zealand Journal of Statistics*, *44*(1), 63–74.

Thompson, S. K. (2006). Adaptive web sampling. *Biometrics*, *62*, 1224–1234.

Thompson, S. K., & Collins, L. M. (2002) Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, *68*, S57–S67.

Thompson, S. K., & Seber, G. A. F. (1996). *Adaptive sampling.* New York: Wiley.

Add-a-Digit Sampling

Add-a-digit sampling is a method of creating a sample of telephone numbers to reach the general public within some geopolitical area of interest. This method is related to directory sampling in that the first step involves drawing a random sample of residential directory-listed telephone numbers from a telephone directory that covers the geographic area of interest. In add-a-digit sampling, the selected directory-listed telephone numbers are not called. Rather, they form the seeds for the list of numbers that will be called. For each directory-listed telephone number drawn from the telephone directory, the last digit of the telephone number is modified by adding one to the last digit. The resulting number is treated as one of the telephone numbers to be sampled. This is the simplest form of add-a-digit sampling. When it was originally devised in the 1970s, it was an important advancement over directory-listed sampling in that the resulting sample of telephone numbers included not only listed numbers but also some numbers that were unlisted residential telephone numbers.

Another practice is to take a seed number and generate several sample telephone numbers by adding 1, 2, 3, 4, 5, and so on to the last digit of the telephone number. However, in the application of this technique, it was found that the higher the value of the digit added to the last digit of the seed telephone number, the less likely the resulting telephone number would be a residential number. Still another method involves drawing the seed telephone numbers and replacing the last two digits with a two-digit random number.

Add-a-digit sampling originated as a method for including residential telephone numbers that are not listed in the telephone directory in the sample. These unlisted numbers are given a zero probability of selection in a directory-listed sample. In add-a-digit sampling, some unlisted telephone numbers will be included in the sample, but it is generally not possible to establish that all unlisted residential telephone numbers have a nonzero probability of selection. Moreover, it is difficult to determine the selection probability of each telephone number in the population, because the listed and unlisted telephone numbers may exhibit different distributions in the population of telephone numbers. For example, one might encounter 500 consecutive telephone numbers that are all unlisted numbers. Because of these and other limitations, add-a-digit sampling is rarely used today. It has been replaced by list-assisted random-digit dialing.

Michael P. Battaglia

See also Directory Sampling; Random-Digit Dialing (RDD); Telephone Surveys

Further Readings

Landon, E. L., & Banks, S. K. (1977). Relative efficiency and bias of plus-one telephone sampling. *Journal of Marketing Research*, *14*, 294–299.

Lavrakas, P. J. (1993).*Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

ADDRESS-BASED SAMPLING

Address-based sampling (ABS) involves the selection of a random sample of addresses from a frame listing of residential addresses. The technique was developed in response to concerns about random-digit dialed (RDD) telephone surveys conducted in the United States because of declining landline frame coverage brought on by an increase in cell phone only households and diminishing geographic specificity as a result of telephone number portability. The development and maintenance of large, computerized address databases can provide researchers with a relatively inexpensive alternative to RDD for drawing household samples. In the United States, address files made available by the U.S. Postal Service (USPS) contain all delivery addresses serviced by the USPS, with the exception of general delivery. Each delivery point is a separate record that conforms to all USPS addressing standards, making the files easy to work with for sampling purposes.

Initial evaluations of the USPS address frame focused on using the information to reduce the costs associated with enumeration of primarily urban households in area probability surveys or in replacing traditional counting and listing methods altogether. These studies showed that for a survey of the general population, the USPS address frame offers coverage of approximately 97% of U.S. households. The frame's standardized format also facilitates geocoding of addresses and linkage to other external data sources, such as the U.S. Census Zip Code Tabulation Areas data. These data can be used to stratify the frame for sampling target populations.

Use of ABS in conjunction with the USPS address frame does have some drawbacks. Researchers cannot obtain the address frame directly from the USPS but must purchase the information through private list vendors. The quality and completeness of the address information obtained from these vendors can vary significantly based on (a) how frequently the company updates the listings, (b) the degree to which the listings are augmented with information from other available databases, and (c) if the company purges records based on requests from householders not to release their information. Moreover, vendors differ in their experience with and ability to draw probability samples from the USPS list. This can be problematic for researchers who do not wish to draw their own samples and tend to rely upon vendor expertise for this task.

Another drawback is that coverage in rural areas tends to be somewhat lower than in urban areas. Additionally, in some rural areas, the USPS files contain simplified (i.e., city, state, and zip code only) listings rather than full street addresses. The percentage of these types of addresses in the database is declining, however, as local governments adopt emergency 911 protocols, which require that all households be identified with a street address. Therefore, over time, simplified address designations are expected to be replaced by full street address information. Another potential issue is that the USPS address frame contains post office (P.O.) boxes and multi-drop addresses (i.e., multiple persons associated with the same address), which may be problematic for both in-person and telephone surveys in which a street address is required to locate the household or to identify a telephone number associated with the household. Such addresses may be less problematic for mail surveys, where the initial goal is to ensure that the mailed questionnaire is delivered to the sampled household.

Households with multiple mailing addresses (e.g., a street address and a residential P.O. box) may also induce selection multiplicities. Research suggests that in some localities a fairly large percentage of households with residential P.O. boxes may also have mail delivered to a street address. Inclusion of P.O. boxes may be necessary, however, to ensure coverage of all households.

Some of the first tests of ABS as an alternative to RDD for general population surveys were conducted by the Centers for Disease Control and Prevention for use on the Behavioral Risk Factor Surveillance System (BRFSS), a large RDD health survey. Two rounds of testing during 2005 and 2006 were conducted with households sampled from the USPS address frame, first using mail surveys, then later utilizing mail surveys with telephone survey follow-up of nonrespondents (a mixed-mode approach). In both instances, the mail survey and mixed-mode approaches produced significantly higher response rates than those obtained in the RDD surveys in states where the RDD response rate was below 40%. The ABS approach also provided access to households with only cell phones, and to a smaller degree, to households with no telephone coverage in percentages that corresponded with other national estimates for the proportional size of these groups. Moreover, the mail survey cost less to conduct than the RDD survey; the mixed-mode approach was cost neutral.

While ABS appears to be an effective sampling frame for conducting mail surveys of the general population, its true potential may be in facilitating mixed-mode surveys. Cross-referencing USPS addresses with other public databases yields telephone numbers for half to two thirds of the addresses. Moreover, ABS may facilitate use of other more cost-effective data collection modes, such as Internet or Web surveys or interactive voice response (IVR). Households could be sampled through ABS, then provided a link to a Web site, given the telephone number for an IVR survey, mailed a hard-copy questionnaire, or any combination of these approaches. Resources permitting, face-to-face surveys could also be added to this mix, particularly since use of the USPS address frame was initially tested as a means of identifying households for such surveys. ABS has the potential, therefore, to serve as a sampling base for a wide variety of single or multi-mode survey designs.

Michael W. Link

See also Area Probability Sample; Cell Phone Only Household; Multi-Stage Sample; Number Portability; Random-Digit Dialing (RDD)

Further Readings

Iannacchione, V. G., Staab, J. M., & Redden, D. T. (2003). Evaluating the use of residential mailing addresses in a metropolitan household survey. *Public Opinion Quarterly*, *76*, 202–210.

Link, M., Battaglia, M., Frankel, M., Osborn, L., & Mokdad, A. (2006). Address-based versus random-digit dialed surveys: Comparison of key health and risk indicators. *American Journal of Epidemiology*, *164*, 1019–1025.

Link, M., Battaglia, M., Frankel, M., Osborn, L., & Mokdad, A. (2008). Comparison of address-based sampling versus random-digit dialing for general population surveys. *Public Opinion Quarterly*, *72*, 6–27.

Advance Contact

Advance contact is any communication to a sampled respondent prior to requesting cooperation and/or presenting the respondent with the actual survey task in order to raise the likelihood (i.e., increase the response propensity) of the potential respondent cooperating with the survey. As explained by Leverage-Saliency Theory, a respondent's decision to participate in research is influenced by several factors, including his or her knowledge of and interest in the survey research topic and/or the survey's sponsor. A researcher can improve the likelihood of a respondent agreeing to participate through efforts to better inform the respondent about the research topic and sponsor through the use of advance contact. Factors in considering the use of advance contacts are (a) the goals of the advance contact and (b) the mode of contact.

The goals of advance contact should be to educate and motivate the respondent to the survey topic and the sponsor in order to improve the likelihood of cooperation with the survey task. The cost and additional effort of advance contact should be balanced against the cost effects of reducing the need for refusal conversion and lessening nonresponse. The first goal of educating respondents is to help them better understand or identify with the topic and/or the sponsor of the research through increasing awareness and positive attitudes toward both. Respondents are more likely to participate when they identify with the research topic or sponsor. Additionally, it is an opportunity to inform the respondent of survey dates, modes of survey participation (e.g., "Watch your U.S. mail for our questionnaire that will be arriving in a first-class [*color and size description of mailer*] around [*anticipated arrival date*]"), and contact information to answer questions or concerns (e.g., "Feel free to contact us toll-free at [*contact number*] or via the Web at [*Web site address*]"). The second goal is to motivate the respondent to participate in the research. This can be done through persuasive messages and appeals to the respondent, such as "Please participate so that your views are represented and represent your community," "This research will help direct money to health care programs in your area," and "This is your chance to make a difference," and so on. Additionally, advance contact is an opportunity to offer or mention incentives (if offered) that the respondent will receive. Research has shown significant improvements in response rate by combining noncontingent cash incentives with advance contact, though the researcher must balance this with research cost and impact to sample representation.

Once the goals of the advance contact have been established, the mode(s) of contact should be selected. The research may select from one or a combination of direct (mail, phone, and email) and indirect (paid advertising, community partnerships, and promotions or special events) modes of advance contact.

A direct mode of advance contact can be via mail or email. A mailed letter or postcard or email (if such an address is available, e.g., when sampling from a membership list) can be used prior to the actual questionnaire being sent or administered to the respondent. Advance mailing can also be a series of contacts that take the form of promotional brochures or flyers that highlight different aspects of the research and/or sponsor. An example used by Nielsen Media Research is the use of mailed brochures highlighting the measurement of the size of the audience for "great moments in television history" (e.g., the first appearance of the Beatles on *The Ed Sullivan Show*) prior to a request to participate in a television viewing survey. Although not used often, a "warm-up" telephone contact (including leaving answering machine messages) also can be used for advance contact.

An indirect mode of advance contact takes the approach of a marketing or public awareness campaign using various forms of communication, including paid advertising in the mass media, community partnerships, and promotions and special community events. Paid (or donated) advertising media can take the form of location-specific media (e.g., billboards, bus or train shelters and benches, flyers) and print and electronic mass media (Internet, magazine, newspaper, radio, and television) such as a public service announcement. Researchers can utilize community partnerships with neighborhood associations or clubs, churches, synagogues, schools, and so on and use a word-of-mouth campaign to spread awareness of research and gain the sponsorship or approval of community leaders. Finally, advance contact can take the form of promotions and special events, such as a booth at a community fair or festival.

Charles D. Shuttles

See also Advance Letter; Fallback Statements; Leverage-Saliency Theory; Nonresponse; Response Propensity; Total Design Method (TDM)

Further Readings

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: Wiley.

Edwards, W. S., & Wilson, M. J. (2003). *Evaluations of the Census 2000 Partnership and Marketing Program.* Retrieved November 3, 2007, from http://www.census.gov/pred/www/eval_top_rpts.htm#PMP

Melgar, A. P., Lavrakas, P. J., & Tompson, T. N. (2004, May). *Evaluating an ad campaign to raise a survey organization's name recognition.* Paper delivered at 59th annual conference of the American Association of Public Opinion Research, Phoenix, AZ.

Shuttles, C. D., Lavrakas, P. J., & Lai, J. W. (2004, May). *Getting respondents to stay on the phone: A large-scale experiment of combining an incentive with a pre-contact letter.* Paper delivered at 59th annual conference of the American Association of Public Opinion Research, Phoenix, AZ.

ADVANCE LETTER

Advance letters (sometimes referred to as "prenotification" letters) are a means of providing potential respondents with positive and timely notice of an impending survey request. The letters often address issues related to the purpose, topic, and sponsor of the survey and a confidentiality promise. In some surveys, advance letters include a token cash incentive. Letters should be sent by first-class mail and timed to arrive only days to a week ahead of the actual survey contact. They also may be accompanied by other informational materials, such as study-related pamphlets, which are typically designed to address questions about survey participation frequently asked by respondents and, in the case of ongoing or longitudinal surveys, provide highlighted results from previous administrations of the survey.

Long used in survey research efforts, advance letters require only that a mailable address be associated with the sampled unit, regardless of whether that unit is a dwelling, telephone number, or name on a listing. Advance letters are used in conjunction with nearly all survey modes, including face-to-face, telephone, mail, and some Web-based surveys. For example, with random-digit dialed (RDD) telephone surveys, sampled telephone numbers are often cross-referenced with electronic telephone directories and other commercially available databases to identify addresses. In a typical RDD sample, addresses can usually be identified for 50–60% of the eligible telephone numbers. Unfortunately, advance letters cannot be used with survey designs when an identifiable address cannot be determined, such as when respondents in the United States are sampled from a frame of cellular telephone numbers or email addresses. Typically, such frames do not include mailable address information.

In terms of content, most of the research literature and best practice recommendations suggest that an advance letter be brief, straightforward, simple, and honest, providing general information about the survey topic without too much detail, especially if the topic is sensitive. The letter should build anticipation rather than provide details or conditions for participation in the survey. Highlighting government sponsorship (e.g., state), emphasizing confidentiality of the data, expressing advance appreciation, and supplying a toll-free telephone number are typically seen as desirable features. Advance letters can also be used to adjust a variety of other influences known to affect survey participation, including use of official stationery of the sponsoring organization to convey legitimacy; having the letter signed by a person in authority; personalizing the name (when available) and address of the sample household and salutation of the letter to convey the importance of the survey; and providing basic information about the nature of the survey questionnaire to educate the household with regard to the task being requested. Additionally, by alerting a household in advance to an upcoming survey request, the letter can be consistent with the norms of politeness that most unannounced contacts from "salespersons" (or even criminals or scam artists) often violate. Furthermore, advance letters can have a positive effect on the interviewers conducting surveys, enhancing their own confidence in seeking a household's participation in a survey.

Postcards are sometimes used in place of actual letters and are considerably less expensive to produce. They also appear, however, less formal and "official" than a letter might; they are more difficult to personalize; they can include less information about the survey than might be included in a letter; and no incentive can be sent with them (nor should one even be mentioned).

Some researchers have argued that it takes only a few seconds to look at a postcard, flip it over, and lay it aside—too short a time for the information to register in the respondent's long-term memory. In addition to being able to enhance a letter over a postcard with more visual and trust-inducing elements, a letter's envelope has to be opened, the letter extracted, reviewed, and then posted, stored, or disposed of, thus increasing the likelihood of the householder's registering it in long-term memory.

Effectiveness and Cost

The effectiveness of advance letters varies with such factors as the length of the letter, the organization on the letterhead, the time lag between mailing and survey contact, and the person to whom the letter is addressed. Particularly germane to the last point, studies indicate that, in about half of households, all the mail is sorted by a single individual, and that 60% discard some mail without opening it, but that this rarely happens to letters addressed to specific household members. Advance letters tend, therefore, to be less effective if their length dissuades people from reading them, if they are not opened and read, if they are read too long prior to contact to recall, and if their sponsorship discounts the value of what is read.

Advance letters can also be accompanied by an incentive (monetary or nonmonetary) to further encourage survey participation. Prepaid cash incentives tend to have the greatest impact on survey participation. Letters can be used, however, to offer a promised incentive, that is, one that is to be provided after completion of a specified task, such as completing an interview. If a noncontingent (pre-paid) incentive is sent in the advance letter, its value should be less than the value of any incentive that is used later in the survey. Past research shows that even $1 or $2 sent in an advance letter will markedly increase the cooperation rate when actual survey contact is made.

The promise of advance letters is that they can increase survey participation, conversely reducing the potential size of nonresponse-related total survey error. For instance, when used in conjunction with RDD telephone surveys, advance letters often have been found to increase response rates by at least 5 percentage points and some times by twice that much. Advance letters can, however, have a heterogeneous impact on subgroups, disproportionately raising participation rates among some groups but not others. This is a problem with many of the techniques developed to reduce nonresponse, particularly those that focus on or are applicable only with a subset of sample members. For instance, in the case of RDD surveys, advance letters can only be used with the subset of respondents for whom an address can be identified; these are disproportionately those respondents who are more likely than average to cooperate in the first place. Likewise, studies have shown that some subgroups are less likely to remember seeing an advance letter sent to their home, in particular, racial minorities, those ages 18 to 34, and those in households with three or more adults. Because survey bias is a function of both the level of nonresponse and the differences between respondents and nonrespondents on measures of importance to the particular survey, improving response rates alone is not enough to guarantee improvement in data quality. Case in point: if efforts to improve participation levels actually exacerbate the distinctions between those who tend to participate in a survey and those who do not, the gains in data quality from reducing nonresponse could actually be offset (or worse, overtaken) by a widening gap between participants and nonparticipants. Researchers should focus, therefore, on reducing overall nonresponse error rather than on simply raising response rates.

In terms of costs, advance letters have been shown in some instances to "pay for themselves." Some studies have shown that the differential cost of obtaining a fixed number of completed interviews from address-matched samples was more than twice as high when advance letters were *not* used, compared to when they were used. In an era of declining survey participation, the fact that this nonresponse-reducing technique often is cost neutral (or nearly so) is welcomed by researchers who are increasingly under pressure to minimize survey costs.

A final consideration: it is impossible to state with certainty that this technique would be effective in reducing nonresponse error in all survey contexts. Researchers are encouraged, therefore, to evaluate the use of advance letters thoroughly within their particular research context to determine whether the gains from the reduction of nonresponse error outweigh the costs or potential for survey bias.

Michael Link

See also Advance Contact; Incentives; Nonresponse; Nonresponse Error

Further Readings

Camburn, D., Lavrakas, P. J., Battaglia, M. P., Massey, J. T., & Wright, R. A. (1996). Using advance respondent letters in random-digit-dialing telephone surveys. *American Statistical Association 1995 Proceedings: Section on Survey Research Methods*, 969–974.

Goldstein, K., & Jennings, M. (2002). The effect of advance letters on cooperation in a list sample telephone survey. *Public Opinion Quarterly*, *66*, 608–17.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Hembroff, L. A., Rusz, D., Ehrlich, N., Rafferty, A., McGee, H. (2004, February). *The cost-effectiveness of alternative advance mailings in a RDD survey*. Paper presented at the Behavioral Risk Factor Surveillance System Annual Conference, Phoenix, AZ.

Link, M., & Mokdad, A. (2005). Use of prenotification letters: An assessment of benefits, costs, and data quality. *Public Opinion Quarterly*, *69*, 572–587.

Agenda Setting

Agenda setting refers to the media effects processes that lead to what are perceived as the most important problems and issues facing a society. It is an important component of public opinion, and thus measuring it accurately is important to public policy deliberation and formation and to public opinion research.

The power to set the public agenda—determining the most important problems for discussion and action—is an essential part of any democratic system. This is so because agenda control is a fundamental lever of power and it is necessary to achieve citizen desires. If democracy is to be a meaningful concept, it must include the right of citizens to have their preferred agenda of topics taken up by policymakers. Leaders who ignore the topics that citizens consider important are not representing the people adequately.

Concepts

Popularized in the mass communication and public opinion literature, agenda setting has for many years been nearly synonymous with studying public issues in a public opinion context. In the study of public opinion, *agenda setting* refers to a type of media effect that occurs when the priorities of the media come to be the priorities of the public. Broadly speaking, the agenda-setting process has three parts:

1. *Public agenda setting* examines the link between issues portrayed in the mass media and the issue priorities of the public.
2. *Policy agenda setting* studies are those examining the activities of public officials or legislatures, and sometimes the link between them and media content.
3. *Media agenda setting* examines the antecedents of media content that relate to issue definition, selection, and emphasis. This can typically include the individual and organizational factors that influence decision making in newsrooms and media organizations generally.

Agenda setting deals fundamentally with the importance or salience of public issues as measured in the popular public opinion polls. Issues are defined similarly to what the polls measure—the economy, trust in government, the environment, and so on—and this ensures comparability to the polling data. The innovation of conceptualizing all the complexity and controversy of a public issue in an abstract manner makes it possible to study issues over long periods of time. But it also tends to produce studies that are quite removed from the very things that made the issues controversial and interesting. Removing details also removes most conflict from the issue. What is left is really just the topic or shell of the issue, with very little content.

Most of the early agenda-setting research focused on the correspondence of aggregate media data and aggregated public opinion data. The rank-order correlations among the two sets of agendas measured the agenda-setting effect. This trend continues to the present day. While it is important to try to understand the connections between media and social priorities, agenda-setting research as it is presently constituted does not do a very good job of explaining how social priorities are really determined. This is so because most agenda-setting research focuses on media as the prime mover in the process and not on the factors that influence the production of media content. Real-world cues are for the most part absent from most agenda-setting studies. Fortunately, new techniques in the analysis of survey data can help revitalize this research tradition. For example, it is becoming easier now to add the respondent's geographical location to survey data. Once one knows the respondent's location, it is possible to append a variety of corresponding contextual or community-level data such as local unemployment rates, taxation levels, housing prices, neighborhood crime rates, and so on. Such contextual data analyzed along with public opinion data using multi-level modeling can help make agenda-setting studies more realistic and inclusive of real-world variables that affect public opinion. Local information about media markets and newspaper circulation areas can also be used in the same way. The key point is that it is important in analysis of agenda-setting effects to make certain that media attention to the problem—and not background conditions—is the real cause.

Background

A famous case study of agenda setting that was developed by Christopher Bosso illustrates this concern with identifying the correct independent and control variables in agenda-setting research. In the case of the Ethiopian famine in 1984, the problem had been at a severe level for some time. Some BBC journalists traveling in Africa filmed sympathetic reports of starving Ethiopians and interested a major American television network in them because of the personal interest of one news anchor. American television news aired the British footage and attracted tremendous interest and more coverage by the other networks and eventually the world. The Ethiopian famine became the subject of worldwide headlines and media attention, from which followed a number of very high-profile food relief efforts and other innovations in fundraising in a global attempt to solve the problem. Of course, the problem had existed long before the media spotlight focused on the problem and continued long after the media tired of the story and moved on. While the audience might conclude that the problem was solved, it was not. But the abrupt spike of interest, as measured by public opinion polls, and subsequent decline and its lack of correlation with the real-world conditions is a classic example of media agenda setting as a unique force, operating by its own logic and according to its own principles. In this case, media acted as a giant searchlight, highlighting an issue for a while, creating considerable interest, and then growing bored of the story and moving on to new problems. The attention of the public often follows. In this case, real-world conditions were not sufficient to explain the public agenda. In fact, the problem is incomprehensible without understanding the media processes.

Political scientist Anthony Downs described this process as the "issue-attention cycle." This model describes a series of stages that certain kinds of long-term chronic problems may go through. The process begins with a pre-problem stage in which the issue exists and experts are aware of it but it has not had much media attention. In stage 2, there is an "alarmed discovery" of the problem accompanied by intense optimism about solving the problem once and for all. This optimism cools considerably by stage 3, in which the true dimensions and costs of the problem become well understood by the public, particularly the nature of the trade-offs and sacrifices that would be required. As Downs explained, a majority of people are likely benefiting from existing conditions and may feel threatened by the kind of fundamental changes that might be needed to overcome many long-standing issues. In the fourth stage there is a general decline of public interest in the problem, accompanied by feelings of discouragement, fear, or boredom. The issue finally settles into a kind of permanent post-problem fifth stage, in which public interest stabilizes at a level well below the peak interest period but higher than it was at the beginning of the cycle. According to Downs's account of the process, sometimes issues stabilize at a level higher than the previous pre-problem stage, but they typically do not regain center stage again for any prolonged period of time.

Not all types of issues are suitable for the cycle of attention described by Downs. Issues likely to receive this type of treatment are those that do not affect the majority of people. The problem is typically caused by power or status arrangements that provide benefits to the majority of people. The final characteristic is that the problem has little or no inherently exciting qualities. In other words, many common social problems such as poverty, racism, transportation, crime, addiction, and unemployment are candidates for this treatment.

As late as the 1980s, the agenda-setting model in mass communication largely meant empirical generalizations based on survey data and content analysis and a set of process variables that included "need for orientation," time lags, topic interest, and media exposure. In the late 1980s, an innovative research program by political psychologists Shanto Iyengar and Donald Kinder used cognitive concepts to reinvent the agenda-setting model, primarily relying mainly on careful experimental methods, although some of their evidence also involved survey data. This work put the agenda-setting model on a firm theoretical footing grounded in social cognitive theory. This led the way to substantial innovation in process terms, as well as work on media priming and media framing, emphasizing different aspects of public issues and the ways they are discussed in public discourse and understood by the public.

In recent years, Maxwell McCombs and his students have continued to develop the agenda-setting model, primarily through efforts to extend the original conceptualization and methods to what they call "second-level agenda setting" or sometimes "attribute agenda setting." This extension of the McCombs agenda-setting tradition attempts to fold the work of media priming and elements of issue framing into his original version of agenda setting. Theoretical benefits of such a project are unclear.

A final consideration is the impact of new media and personalized systems of communication on the future of agenda setting. This is an important consideration, because agenda setting dates from the mass communication era. One distinctive feature of the mass communication system during the past decade has been the proliferation of channels through which news flows and that audiences use to become informed. The rich variety of outlets, including multiple channels of cable and satellite television, newspapers, and online sources, makes studying the news agenda no longer the simple process that it used to be. In his original 1972 study, McCombs could analyze the newspaper reports in one city and represent the media agenda to which that community had been exposed. This is impossible today, given the wide range of available communication outlets. In addition to increased diversity of channels of communication, a person's media use can be readily customized to an unprecedented degree.

Looking Forward

Studying agenda setting in the new information environment where "Search Is Everything" will be increasingly challenging. One way to address this issue is to focus more research attention on the political economy of search engines that are delivering news to many people and the agenda-setting power of their methods to determine who sees what news. Search engines operate via proprietary algorithms that they apply to the portion of the Internet that they are able to map and index. When a user enters a topic into a search engine, the search engine returns a prioritized list—an agenda—of results. Unfortunately, how this agenda is set is anything but transparent. In fact, search results vary, sometimes dramatically, from search engine to search engine based on the nature of the formulae used to find the results and prioritize them. Most search engines collect fees from clients who want their search terms to appear higher on the prioritized order of results. Some disclose that a given site's result is a "sponsored link," but this is not a universal practice. In other words, commercial interests often buy the answer to a given search. Search results can also be influenced without anyone making a payment directly to a search engine. Results are "gamed" by firms known as *optimizers*, which collect fees in exchange for figuring out ways to move certain results higher on the list. They do this through painstaking attempts to learn key elements of the algorithms used to determine the agenda order and then making sure their clients' sites meet these criteria.

In an information environment that increasingly depends on search technology, the political economy of search is an understudied but key component of what the public knows and thinks is important: the public agenda. In today's fracturing media environment, consumers and citizens rely increasingly on standing orders for customized information that meets certain specifications. How that information is searched and delivered will be an increasingly significant issue for political and commercial interests as well as public opinion researchers seeking to understand the public's priorities. A challenge to survey researchers will be to understand this process and use it to design studies that incorporate an up-to-date understanding of the media system. This can help assure the relevance of the agenda-setting model for years to come.

Gerald M. Kosicki

See also Issue Definition (Framing); Multi-Level Integrated Database Approach (MIDA); Priming; Public Opinion; Public Opinion Research

Further Readings

Bosso, C. (1989). Setting the agenda: Mass media and the discovery of famine in Ethiopia. In M. Margolis & G. A. Mauser (Eds.), *Manipulating public opinion: Essays on public opinion as a dependent variable* (pp. 153–174). Pacific Grove, CA: Brooks/Cole.

Dearing, J. W., & Rogers, E. M. (1996). *Communication concepts 6: Agenda-setting*. Thousand Oaks, CA: Sage.

Downs, A. (1972). Up and down with ecology: The issue-attention cycle. *The Public Interest*, *28*, 38–50.

Iyengar, S., & Kinder, D. R. (1987). *News that matters.* Chicago: University of Chicago Press.

Iyengar, S., & Kinder, D. R. (1994). *Is anyone responsible? How television frames political issues.* Chicago: University of Chicago Press.

Kosicki, G. M. (1993). Problems and opportunities in agenda-setting. *Journal of Communication*, *43*, 100–128.

McCombs, M. (2004). *Setting the agenda: The mass media and public opinion.* Cambridge, UK: Polity Press.

McCombs, M., & Shaw, D. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*, 176–187.

Scheufele, D. A. (2000). Agenda-setting, priming and framing revisited: Another look at cognitive effects of political communication. *Mass Communication & Society*, *3*, 297–316.

Aided Recall

Aided recall is a question-asking strategy in which survey respondents are provided with a number of cues to facilitate their memory of particular responses that are of relevance to the purpose of the study. Typically such cues involve asking respondents separate questions that amount to a list of subcategories of some larger phenomenon. The purpose of listing each category and asking about it separately is to assist the respondent by providing cues that will facilitate memory regarding that particular category.

Applications

This question technique is most appropriate when the researcher is most concerned about completeness and accuracy and more worried about underreporting answers than in overreporting. Aided recall question strategies structure the range of possible answers completely and simplify the task for the respondent. They also simplify the investigator's work in gathering and analyzing the data, since no recording or coding of open-ended protocols is required, according to Seymour Sudman and Norman Bradburn in their classic volume, *Asking Questions.*

While it might seem most natural to ask respondents to self-nominate events to be recalled or criteria that they will use in decision making, they may easily forget or overlook relevant answers. This can occur for many reasons. The respondent might not take the time to think the answer through carefully and completely. The respondent might think that certain potential aspects of his or her answer are not relevant or appropriate and so are omitted. Respondents might not want to take the time needed to respond to the questions or could be hurried along by an interviewer. Difficult or time-consuming tasks might encourage respondents to satisfice—that is, to report what comes to mind as the first acceptable answer or use other mental shortcuts—rather than optimizing their answers by making them as complete and thoughtful as possible. When forgetting seems particularly likely, aided recall questions should be used, as recommended by Sudman and Bradburn.

Aided recall questions are common in the survey literature. An example will help to clarify the strategy, as will a contrast to *unaided* recall. To ask respondents about where they typically obtain public affairs information, one might simply ask a broad, open-ended question and attempt to code the responses until the respondent had been thoroughly probed and had nothing else to say. This would be an example of unaided recall. The respondent would be given no clues to limit or steer the scope of the inquiry and would have to conduct a thorough information search of his or her own memory to think of possible answers as well as to screen them in terms of appropriateness. If the respondent answered by mentioning radio, television, and newspapers, the interviewer might probe further by asking if there were any other sources. Uncertain of how detailed to make the answer, at that time the respondent might mention magazines. The person might not have thought that online sources of information were appropriate or may simply not think of them at the time. Another possibility is that an additional interviewer probe might have elicited online sources.

A variation on this general topic domain using an aided recall strategy might ask about what sources the respondent used for public affairs information in the past week and then might proceed to list a number of such sources. By listing each source explicitly and asking whether or not the respondent used it, the survey designer is enhancing completeness and prompting the respondent to think of the meaning of the topic in the same way. In this way there is less opportunity for the respondent to overlook possible categories, but he or she may feel under more pressure to agree to more categories for fear of appearing uninformed. Sources that might be mentioned in the answer include daily and weekly newspapers, news magazines, local and national on-air television, cable-only television networks such as CNN, CNBC, and FOX, and the various channels of C-SPAN. They might also include various popular online sources of news such as Yahoo.com, MSN.com, Google News, and *The New York Times* Web site, as well as interpersonal channels of communication such as friends, coworkers, and family members. In addition to all of these clearly specified information channels, one should also probe for other responses not listed.

Simpler variations on aided recall include listing some examples of the kind of general responses that are anticipated or showing respondents a card containing a list of possible responses and asking them to indicate which ones apply to their situation. This information ensures that respondents do not forget to consider items of particular importance to the purposes of the question. To ensure the meaningfulness of such questions, the list of items from which respondents choose must be complete. Such completeness can be guided by theoretical concerns and literature and verified by pretesting. Such questions can only be as valid as the completeness of the list. The order in which items on the list are presented to the respondents also is an important issue; ideally this should be varied systematically or randomly across respondents. Very long lists should be avoided, as they can make respondents feel that they need to respond positively to at least some of the items. Sudman and Bradburn suggest that when lists become long, questionnaire designers should consider a system of screening questions.

In general, the aided recall question strategy will yield higher estimates of what is measured compared to unaided recall items. However, the list tends to convey to the respondent at least implicit expectations for positive responses to something on the list. While aided recall questions are helpful when underreporting is likely to be an issue, they can lead to overreporting. They are thus inappropriate in situations in which overreporting is likely to be a problem, or at least they need to be used with other tools that will help limit overreporting, such as screening questions. Roger Tourangeau, Lance Rips, and Ken Rasinski's book, *The Psychology of the Survey Response*, provides extensive relevant discussions of the theoretical issues related to these problems of memory and the survey response.

Gerald M. Kosicki

See also Aided Recognition; Cognitive Aspects of Survey Methodology (CASM); Satisficing; Show Card; Unaided Recall

Further Readings

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York: Cambridge University Press.

Aided Recognition

Within the context of survey research, aided recognition is a form of aided recall in which a survey respondent is asked if she or he was aware of something prior to being asked about it in the survey questionnaire. The stimulus that the respondent is asked about typically is the name of a company or of a product or service. In some cases, other than in telephone surveys, a picture can be shown as the stimulus. In telephone, Internet, and in-person surveys, audio can serve as the stimulus for the respondent.

The common form for measuring aided recognition is to use a closed-ended survey question along the following lines:

> Before today, have you ever heard of _____?

The respondent is asked to simply answer "Yes" or "No." Sometimes a respondent is uncertain and says so to the interviewer. Thus the questionnaire can be precoded with an "Uncertain/Maybe/etc." response that is not read to the respondent but that an interviewer can code if the respondent volunteers such.

Aided recognition is often used in branding studies as a measure of people's awareness of a company brand. Typically this is done by mixing the name of the brand that is the primary focus of the survey with the names of competitors in series of separate items. In this way, the survey can show how recognition levels compare across brands. It often is prudent to include at least one "bogus" brand name in the list of brands asked about to measure the baseline level of "Yes" saying among respondents, which is a form of acquiescence response bias. If a series of aided recognition items is asked, it also is prudent to use either a random start or a random order in presenting the items in the series to different respondents.

Aided recognition questions must be asked *after* any unaided recall questions are asked on the same topic; otherwise the aided recognition questions will bias answers to the unaided recall questions. Subsequent to the positioning of unaided recall and aided recognition questions within a questionnaire, branding studies often include image questions about the brand to get more information on the valence (positive or negative) associated with the brand. Logic dictates that any respondent who is not able to mention the brand under the unaided recall questions or to recognize the brand under the aided recognition questions is not asked any of the image questions.

Paul J. Lavrakas

See also Acquiescence Response Bias; Aided Recall; Bogus Question; Closed-Ended Question; Precoded Question; Random Order; Random Start; Unaided Recall

Further Readings

Eastman, S. T. (2000). *Research in media promotion.* Mahwah, NJ: Lawrence Erlbaum.

Algorithm

Algorithm is a computer science term for a way of solving a problem, and it also refers to the instructions given to the computer to solve the problem. The study of algorithms is central to computer science and is of great practical importance to survey data analysis because algorithms are used in statistical programs.

An algorithm can be thought of as any step-by-step procedure for solving a task. Imagine five playing cards face down on a table and the task of sorting them. Picking them up one at a time with the right hand and placing them in the left hand in their proper

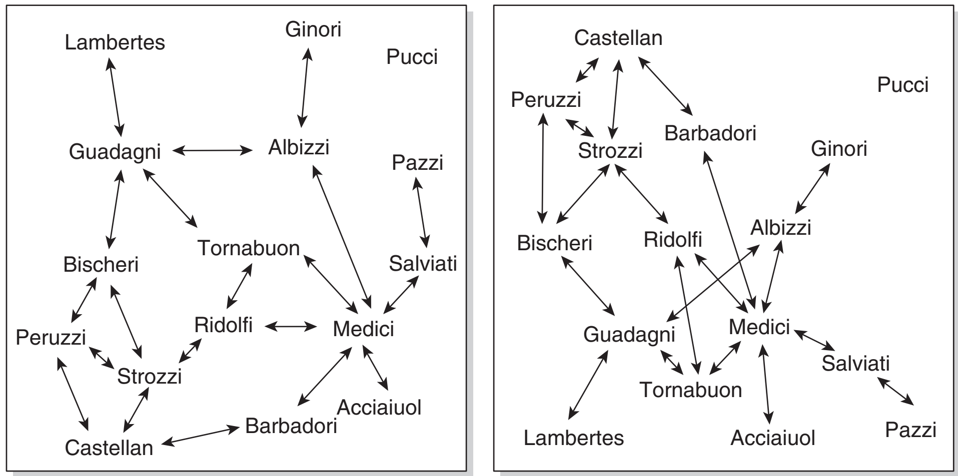

Figure 1 Two possible depictions of the same network data

place would be one way to solve this task. This is an algorithm, called *insertion sort* in computer science.

It is worth noting the subtle distinction between the concept of algorithm and the concept of a *method* or of a *technique.* For example, a method would be least squares; matrix inversion would be a technique used therein; and LU decomposition and Strassen's algorithm would be alternative algorithms to accomplish matrix inversion. A single data analysis method may use more than one algorithm.

It is impossible to write statistical software without using algorithms, so the importance of algorithms to survey data analysis is assured. However, user-friendly statistical software packages eliminate the need for end users to construct their own algorithms for most tasks. Nonetheless, at least a basic understanding of algorithms can be useful to survey researchers. For example, maximum likelihood methods can use an initial estimate as a starting point, and in some cases failure to converge may be remediated by trivially altering the initial estimate. Without some familiarity of the underlying algorithm, a researcher may be stuck with a nonconverging function.

Another setting where some knowledge of algorithms is useful is shown in Figure 1, which illustrates two possible depictions of the exact same network data. The left panel uses the multi-dimensional scaling algorithm and the right uses simulated annealing. The data are identical, which may be verified by observing who is connected to whom, but the appearance of the graphs is different. Algorithms are important here, because interpretation of the network data is affected by the appearance of the graph, which is affected in turn by the choice of algorithm. Whereas in many cases different algorithms will produce the same result but differ in speed (i.e., computing time), in this case different algorithms produce different results.

The term *algorithm* is sometimes used more broadly to mean any step-by-step procedure to solve a given task, whether or not a computer is involved. For instance, matching historical records from more than one archival source can be done by hand using an algorithm. Moreover, it is not only the analysis of survey data that uses algorithms, but also in many cases in the collection of the data an algorithm may be used to select clusters in a complex sample survey design.

Andrew Noymer

Further Readings

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge: MIT Press.

Knuth, D. E. (1997). *Fundamental algorithms: The art of computer programming* (3rd ed., Vol. 1). Reading, MA: Addison-Wesley.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). Cambridge, UK: Cambridge University Press.

Alpha, Significance Level of Test

Alpha is a threshold value used to judge whether a test statistic is statistically significant. It is chosen by the researcher. Alpha represents an acceptable probability of a Type I error in a statistical test. Because alpha corresponds to a probability, it can range from 0 to 1. In practice, 0.01, 0.05, and 0.1 are the most commonly used values for alpha, representing a 1%, 5%, and 10% chance of a Type I error occurring (i.e., rejecting the null hypothesis when it is in fact correct). If the *p*-value of a test is equal to or less than the chosen level of alpha, it is deemed statistically significant; otherwise it is not.

The typical level of alpha is 0.05, but this is simply a custom and is not based on any statistical science theory or criteria other than conventional practice that has become the accepted standard. Alpha levels of 0.1 are sometimes used, which is a more lenient standard; alpha levels greater than 0.1 are rarely if ever used. All things being equal, standard errors will be larger in smaller data sets, so it may make sense to choose 0.1 for alpha in a smaller data set. Similarly, in large data sets (hundreds of thousands of observations or more), it is not uncommon for nearly every test to be significant at the alpha 0.05 level; therefore the more stringent level of 0.01 is often used (or even 0.001 in some instances). In tabular presentation of results, different symbols are often used to denote significance at different values of alpha (e.g., one asterisk for 0.05, two asterisks for 0.01, three asterisks for 0.001). When *p*-values of tests are reported, it is redundant also to state significance at a given alpha.

Best practice is to specify alpha before analyzing data. Specifying alpha after performing an analysis opens one up to the temptation to tailor significance levels to fit the results. For example, if a test has a *p*-value of 0.07, this is not significant at the customary 0.05 level but it meets what sometimes is referred to as "marginal" significance at the 0.1 level. If one chooses a level of alpha after running the model, nothing would prevent, in this example, an investigator from choosing 0.1 simply because it achieves significance. On the other hand, if alpha is specified a priori, then the investigator would have to justify choosing 0.1 as alpha for reasons other than simply "moving the goalposts." Another reason to specify alpha in advance is that sample size calculations require a value for alpha (or for the confidence level, which is just 1 minus alpha).

Note that if 20 statistical models are run, for example, then one should expect one of them to produce a significant result when alpha is set at 0.05, merely by chance. When multiple tests are performed, investigators sometimes use corrections, such as the Bonferroni correction, to adjust for this. In and of itself, specifying a stringent alpha (e.g., 0.01 or 0.001) is not a guarantee of anything. In particular, if a statistical model is misspecified, alpha does not change that.

Only models in which a given alpha is satisfied tend to reach consumers, who tend to be exposed to scientific studies via referred journal articles. This phenomenon is known as "publication bias." The reader of a study may find it persuasive because the *p*-value is smaller than alpha. The persuasion derives from the small likelihood (alpha) of the data having arisen by chance if the null hypothesis is correct (the null hypothesis is therefore rejected). But even at a small level of alpha, any given result may be likely by sheer chance if enough models have been run, whether or not these models are reported to the reader. Even an arbitrarily small alpha is meaningless as a probability-based measure if many models are run and only the successful ones revealed. A small level of alpha, taken by itself, is therefore not an indicator that a given piece of research is persuasive.

Statistical models are sometimes used for purely descriptive purposes, and in such contexts no level of alpha need be specified.

Andrew Noymer

See also Null Hypothesis; Probability; *p*-Value; Standard Error; Type I Error

Further Readings

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *57*, 289–300.

Neyman, J., & Pearson, E. S. (1967). The testing of statistical hypotheses in relation to probabilities a priori. In J. Neyman & E. S. Pearson, *Joint Statistical Papers of J. Neyman and E. S. Pearson* (pp. 186–202). London: Cambridge University Press.

Alternative Hypothesis

An alternative hypothesis is one in which a difference (or an effect) between two or more variables is anticipated by the researchers; that is, the observed pattern of the data is not due to a chance occurrence. This follows from the tenets of science, in which empirical evidence must be found to refute the null hypothesis before one can claim support for an alternative hypothesis (i.e., there is in fact a reliable difference or effect in whatever is being studied). The concept of the alternative hypothesis is a central part of formal hypothesis testing.

Alternative hypotheses can be nondirectional or directional. If nondirectional, an alternative hypothesis is tested with a two-tailed statistical test and is stated in words to the effect that "A differs from B." If directional, an alternative hypothesis is tested with a one-tailed statistical test and is stated in words to the effect that "A is greater than B" or "B is greater than A." (The null hypothesis is stated in words to the effect that "A equals B.")

An example in survey research would be a split-half experiment that is used to test whether the order of two question sequences within a questionnaire affects the answers given to the items in one of the sequences, for example, in crime surveys where both fear of crime and criminal victimization experience are measured. In this example, a researcher could venture a directional alternative hypothesis that greater levels of fear would be reported if the fear items followed the victimization items, compared to if they preceded the victimization items. Half the respondents would be randomly assigned to receive one order (fear items, then victimization items), and the other half would receive the other order (victimization items, then fear items). The null hypothesis would be that the order of these question sequences makes no difference in the answers given to the fear-of-crime items. Thus, if the null hypothesis is true, the researcher would not expect to observe any reliable (i.e., statistically significant) difference in levels of fear reported under the two question-ordering conditions. In contrast, if the directional alternative hypothesis is true (i.e., if results indicate significantly greater fear being reported when the fear items follow the victimization items than when they precede them), then the null hypothesis is rejected and support is accorded to the alternate hypothesis.

Another way of understanding the alternative and null hypotheses in survey research is to think about the crime survey example and the confidence intervals that can be calculated around the fear-of-crime measures in the two conditions. The null hypothesis would be that the 95% confidence intervals for the fear measures under the two question orders would overlap and thus not be reliably (significantly) different from each other at the .05 (alpha) level. A directional alternative hypothesis that states that reported fear of crime would be higher when the victimization items precede the fear items would be that (a) the confidence intervals would not overlap and that (b) the lower limit of the confidence interval for the fear items when the victimization items precede them would exceed the upper limit of the confidence interval for the fear items when the victimization items follow them.

Supporting an alternative hypothesis when it is in fact false is termed a *Type I error*. Failing to support an alternative hypothesis when it is in fact true is termed a *Type II error*.

Paul J. Lavrakas

See also Alpha, Significance Level of Test; Confidence Interval; Experimental Design; Null Hypothesis; *p*-Value; Split-Half; Statistical Power; Type I Error; Type II Error

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth/Cengage Learning.

American Association for Public Opinion Research (AAPOR)

The American Association for Public Opinion Research (AAPOR) is the principal professional association for survey researchers in the United States. Organized shortly after World War II, AAPOR develops and promotes ethical principles to guide survey research, advances its methodology, and attempts to further an understanding of appropriate practice both for researchers and the general public. Its ethical code and its enforcement have evolved with changing technology and new applications of survey research.

Founding of AAPOR

The redeployment of U.S. industrial power to the production of consumer goods after World War II stimulated interest in a wide variety of survey applications, particularly market and media research. The economy needed mass media to sell the output of mass production, and survey research made the marketing process efficient.

Harry Field, who had founded the National Opinion Research Center (NORC) at the University of Denver in 1941, saw the war's end as an opportunity to assemble the diverse strands of survey research. He organized a national conference to open on July 29, 1946. The site was Central City, Colorado, 42 miles of winding mountain road from downtown Denver and 8 hours by reciprocating-engine airliner from New York City. Field invited 264 practitioners, and 73 attended. Don Cahalan, who coordinated the event, classified the attendees: 19 from media, 18 academics, 13 commercial researchers, 11 from nonprofits, 7 government employees, 3 from advertising agencies, and 2 others.

A key session on technical and ethical standards in public opinion research was led by George Gallup, Clyde Hart of the Office of Price Administration, Julian Woodward of Elmo Roper's organization, and Field. In a paper that Paul Sheatsley would later describe as "remarkably prescient," Woodward foresaw expanded use of polls to provide feedback for elected officials and to test public knowledge. Competition among polls would create pressure to minimize costs, but because such polls would play an important role in public service by providing a continuing referendum on policy and consumer issues, they would require standards of quality that would "justify the responsibilities which will increasingly be theirs."

After 3 days of discussion, the conference decided that a second meeting should be held in 1947. Harry Field was to lead it, but he died in a plane crash in France only a month later. Clyde Hart became director of NORC and organizer of the second conference.

For the second meeting, Hart and the sponsoring committee chose Williamstown, Massachusetts, in the northwest corner of the state. Julian Woodward assembled a program that drew 194 participants.

While the Central City meeting had envisioned an international confederation of existing survey research organizations, the Williamstown meeting took the unexpected step of forming a membership organization instead. A constitution was drafted, and the name "American Association for Public Opinion Research" was approved after assurances were made that an international organization would be formed the next day. Since that time, AAPOR and the World Association for Public Information Research (or WAPOR) have combined their meetings in even-numbered years.

Clyde Hart was elected by acclamation, and, in a secret ballot, Elmo Wilson, research director for CBS, was named vice president. Wilson's election as president the following year began the AAPOR tradition of alternating the presidency between the commercial and academic sectors. A 1951 revision of the constitution provided for the vice president to ascend automatically to the presidency.

Mission of AAPOR

One function of a professional association is to codify the profession's self-definition by setting standards of ethics and technical competence. When AAPOR was founded, the main technical debate was between the advocates of quota sampling and those who preferred probability sampling. It quickly became clear that setting rules of scientific orthodoxy was not practical, but there was support for setting moral standards, particularly regarding transparency in research methods.

The other key aspect of professionalism is advancement of the profession's body of knowledge. The constitution adopted at Williamstown provided for the "dissemination of opinion research methods, techniques and findings through annual conferences and an official journal and other publications." *Public Opinion Quarterly* had been started in 1937 at Princeton University, and AAPOR designated it the official journal of the association, paying a fee to have its conference proceedings published there. In 1968, the journal was acquired by Columbia University, and title was transferred to AAPOR in 1985.

Evolution and Application of the AAPOR Code

Several years passed without the association having to face a specific case or controversy. That ended in 1955, when Walter Reuther, president of the United Auto Workers, filed a complaint alleging biased questions in a survey of General Motors employees. The Standards Committee of AAPOR shied away from dealing with the issue and sent a summary of the case to the membership so that "each is free to make his own evaluation."

Sidney Hollander, in his 1992 history of the Standards Committee, found the next critical point to occur in 1957, when members became concerned about a conflict between their duty to maintain the confidentiality of survey respondents and possible demands for their names as legal evidence. Researchers would have a stronger case if respondent anonymity could be specified as a professional standard.

That need opened the door to the development of a formal code. Different versions were presented to the 1958 and 1959 meetings without success; finally a code was adopted at the 1960 annual meeting with responsibility for enforcement assigned to the Executive Council.

The standards became more specific in 1967 with the adoption of disclosure requirements—key pieces of information that should be revealed about any poll, for example, sample size, dates of interviewing, question wording, method of data collection, and identity of the sponsor of the survey. A test case arose in 1974 when survey findings supporting the Nixon administration were released without identifying the sponsor, which turned out to be the Republican National Committee. No action was taken because AAPOR lacked defined procedures for enforcing its rules.

That flaw was repaired under the leadership of California pollster Mervin Field during his tenure as Standards chair in 1974–1975. A detailed procedure was worked out to provide formal hearings, right of reply, and protection of the anonymity of accusers. In its first application, the procedure led to a finding that Opinion Research Corporation, in a survey report used to oppose establishment of a federal consumer advocacy agency, had made interpretations unsupported by the publicly released data.

One effect was to give journalists a tool to extract information from reluctant pollsters. Survey researchers could not hide behind confidentiality obligations to their clients if to do so would conceal a violation of good practice. The code, which every member signs, contains this language: "If we become aware of the appearance in public of serious inaccuracies or distortions regarding our research, we shall publicly disclose what is required to correct these inaccuracies or distortions . . . ."

A person need not be a member of AAPOR to lodge a complaint, nor does AAPOR limit its investigations to members. From 1975 to 1997, the organization used publicity as a sanction in the form of a press release issued after a council finding. The organization fell relatively silent after 1997, continuing to investigate complaints of code violations but imposing sanctions by private letter of censure with no public announcement.

Much of the recent effort at enforcing standards has been directed at pseudo-polls used to cover generation of marketing leads, develop voter lists, or disseminate political falsehoods. The organization also turned its attention to education and promotion, hiring its first full-time public relations specialist in 2007.

Annual AAPOR Conference

The annual conference has traditionally included a plenary session on a current topic of broad interest, an address by the current president, formal paper presentations organized by topic with discussants, round table discussions, teaching sessions, and informal networking. In the early days, conference organizers favored university settings for the sake of economy, but as the organization grew, resort hotels became the standard choice. Further growth, with conference attendance approaching 1,000, drew the meetings to metropolitan areas. By the early 21st century, AAPOR had become an organization of more than 2,000 members with annual revenue of nearly $1 million.

Philip Meyer

See also Anonymity; Confidentiality; Disclosure; Ethical Principles; Gallup, George; National Opinion Research Center (NORC); Probability Sampling; Pseudo-Polls; *Public Opinion Quarterly* (POQ); Quota Sampling; Roper, Elmo; Sheatsley, Paul; World Association for Public Opinion Research (WAPOR)

Further Readings

American Association for Public Opinion Research: http://www.aapor.org

Cahalan, D. (1992). Origins: The central city conference. In P. Sheatsley & W. Mitofsky (Eds.), *A meeting place: The history of the American Association for Public Opinion Research* (pp. 25–40). Lenexa, KS: American Association for Public Opinion Research.

Hollander, S. (1992). Survey standards. In P. Sheatsley & W. Mitofsky (Eds.), *A meeting place: The history of the American Association for Public Opinion Research* (pp. 65–103). Lenexa, KS: American Association for Public Opinion Research.

Phillips Davison, W. (1992). AAPOR and the printed word. In P. Sheatsley & W. Mitofsky (Eds.), *A meeting place: The history of the American Association for Public Opinion Research* (pp. 241–248). Lenexa, KS: American Association for Public Opinion Research.

Sheatsley, P. (1992). The founding of AAPOR. In P. Sheatsley & W. Mitofsky (Eds.), *A meeting place: The history of the*

American Association for Public Opinion Research (pp. 41–62). Lenexa, KS: American Association for Public Opinion Research.

American Community Survey (ACS)

The American Community Survey (ACS) is an ongoing national survey conducted by the U.S. Census Bureau. Part of the federal decennial census program, the ACS was designed to replace the long form or sample portion of the decennial census, starting in 2010. By conducting monthly surveys of a sample of the U.S. population, the ACS collects economic, social, and housing information continuously rather than every 10 years. The ACS does not replace the decennial enumeration, which is constitutionally mandated for apportioning congressional seats. It is expected that the ACS program will improve the quality of the decennial census, because the elimination of long-form questions should increase response and allow more focused nonresponse follow-up.

Eventually, the ACS will supply data for the same geographic levels that have traditionally been available from the census long form, including sub-county areas such as census tracts and block groups. The ACS sample sizes are not large enough to support annual releases for all geographic areas. For smaller areas, the ACS data are averaged over multiple years. Annual data are available for populations of 65,000 or more. Annual estimates from the 2005 ACS were released in 2006. Three-year averages will be released for areas with 20,000 or more people, and 5-year averages will be available for the remaining areas. Three-year averaged data will be available starting in 2008, and the 5-year averaged data will first be available in 2010. After 2010, data for all geographic data will be updated annually, using the rolling 3- or 5-year averages for the smaller areas.

The Census Bureau has conducted ACS tests in select counties since the mid-1990s. In 2005, the housing unit sample was expanded to its full size, which includes all U.S. counties and equivalents, the District of Columbia, and Puerto Rico. The ACS was expanded to include group quarters facilities in 2006. As an ongoing program, funding for the American Community Survey must be approved by Congress annually as part of the federal budget process. Current ACS implementation plans could change in the future if funding is not approved.

Content

Recent versions of the ACS questionnaires have included the same general subjects as the 2000 long form, asking more than 20 housing questions and more than 30 population questions about each household member. The population questions include the six basic demographic questions from the 2000 census short form (name, relationship to householder, age, sex, Hispanic identity, and race). ACS questions cover subjects such as ancestry, language use, education, occupation, veteran status, income, and housing costs. The content remained the same for the 2005 and 2006 surveys and is planned to remain the same for 2007.

The content of the American Community Survey is determined through a formal process managed by the Census Bureau and the federal Office of Management and Budget (OMB). The Census Bureau and OMB restrict ACS content to include only questions that are necessary for a specified federal purpose, such as a regulation that requires use of the subject data. Because the ACS is a continuous survey, changes to the survey can result in inconsistent data trends. Content changes are minimized and cannot be made more than once per year. Content modifications require extensive testing. Census Bureau staff and other subject experts review content test results and make recommendations to the OMB, which makes final content decisions.

Sample Design and Selection

The American Community Survey is stratified so that housing units and group quarters facilities are sampled separately. On average, sample rates for both populations are targeted to be 2.5% per year. Approximately 250,000 housing unit addresses are selected in each month, or 3 million per year. The ACS selects addresses from the Census Bureau's Master Address File (MAF). The MAF is a list of housing units and group quarters facilities in the United States. Because the completeness of the sample frame is so important to the ACS sample process, the MAF file is reviewed and updated on an ongoing basis. To update the MAF, the Census Bureau uses information from the U.S. Postal Service and from local governments.

For each ACS sample year, there are two phases for selecting the addresses. The first phase takes place a few months prior to the sample year, and a supplemental phase takes place early in the sample year. The supplemental phase allows for the inclusion of addresses that have been added since the first phase. The ACS allocates addresses to subframes to ensure that no address can be chosen more than once during a 5-year period.

The ACS intends to provide reliable data for local areas of varying sizes. The ACS staff must also intensely protect the confidentiality of respondents. In order to meet the reliability and confidentiality standards and still report data for very small areas, the Census Bureau employs differential sample rates. In this process, the sample is stratified so that addresses in smaller geographic areas have a higher probability of selection than those in larger areas.

Data Collection and Processing

ACS surveys are administered using three collection modes: mail, telephone, and in person. Addresses that are determined to be incomplete are also assigned for in-person collection. The large majority of households are contacted first through the mail. The mail-out process begins with a pre-survey letter that notifies the recipients that they will receive a survey. Next the complete survey packet is sent, including a cover letter, the questionnaire, instructional guidance, and a return envelope. A reminder postcard is sent to all mail recipients several days after the survey packet. After a number of weeks, if questionnaires are not returned, the Census Bureau will send another survey packet. The ACS typically has maintained very high mail-back response rates.

Respondents who return incomplete surveys or do not mail back surveys after a designated amount of time will be contacted by telephone. Using a computer-assisted telephone interview (CATI) process, Census Bureau interviewers will attempt to complete the survey on the phone.

Surveys that are not completed by mail or telephone will become eligible for in-person data collection through a computer-assisted personal interview process (CAPI). Because of the high costs of in-person data collection and the difficulty in reaching persons who have not responded during other phases, not all of these nonresponse cases will be chosen for personal interview. The ACS selects a subsample of nonrespondents for the CAPI phase. The responses from the nonresponse follow-up are weighted up to account for the nonrespondents who are not contacted.

Currently, standard ACS questionnaires are produced in English and in Spanish. English forms are mailed to homes in the United States, and Spanish forms are mailed to homes in Puerto Rico. ACS questionnaires include phone numbers that recipients can call for assistance in filling out the questionnaire. English forms include these phone assistance instructions in both English and Spanish. Persons in the United States may request the Spanish language form.

Sources of Survey Error in the ACS

A sample-based survey, the ACS will have sampling and nonsampling error. Sampling error is the random error that occurs when the survey is conducted for a sample of the universe rather than for all members of the universe. Sampling errors are often described using standard errors and margins of error. ACS data are published with margins of error at the 90% confidence level.

The ACS is also subject to nonresponse error through both unit and item nonresponse. *Unit nonresponse* occurs when recipients do not return their ACS forms or mail back blank forms. *Item nonresponse* occurs when certain questions are not answered. Compared to other surveys, the ACS has maintained relatively low levels of both unit and item nonresponse. One reason for the high response rates is that, like decennial census, persons who are selected for the ACS are required by law to participate. Another contributing factor to the high response rates relates to fact that the ACS is an ongoing operation. Unlike the decennial census and other less frequent surveys, the ACS maintains a regular staff of professional interviewers who receive in-depth training on how to gain cooperation and collect information during the telephone and in-persons phases.

General ACS Considerations

Users will find that there a number of things to keep in mind when using ACS data, especially when making comparisons to decennial census data. Users need to adjust to the multi-year averages as well as to the higher rates of sampling error. While the 2000 census long form was sent to 1 in 6 housing units, the ACS will be sent to about 1 in 8 households in a 5-year period. Thus, to provide the more frequent data updates, there has been a trade-off in the size of the samples. When comparing data, only statistically significant changes should be considered. The Census Bureau publishes instructions for users on how to apply statistical tests when trying to measure change over time.

Because the ACS is conducted monthly, annual ACS data essentially reflect an average throughout the year. In contrast, the decennial census reflected a particular point in time (traditionally April of the census year). This consideration is particularly important when comparing data for areas that have seasonal population fluctuations, such as college towns or resort areas.

The ACS also employs different residency rules than the decennial census. While the decennial census counts people in their usual place of residence (where they spend the majority of the year), the ACS includes people who have lived in the sample residence for most of the past 2 months.

Questions about concepts such as income and mobility are also conducted differently with the ACS. While the decennial census asks for income amounts for the prior year; the ACS asks for income over the past 12 months. In the 2000 census, respondents were asked if they lived in the housing unit on April 1, 1995. The ACS question asks whether the resident lived in the unit 1 year ago.

The ACS is designed to provide information about the characteristics of U.S. populations, but it is not designed to provide annual updates to the decennial census total population or housing unit counts. The official responsibility for updating population estimates falls under the Census Bureau's Population Division, which produces annual estimates of the total population and population by age, sex, race, and Hispanic identity. The estimates are produced for the nation, states, and for all U.S. counties and county equivalents. To estimate the population, the Census Bureau uses the components-of-change approach, which estimates change from the 2000 decennial census base counts. The components of population change are births, deaths, and migration. To estimate the components of change, the Census Bureau uses sources such as birth records, death certificates, and Internal Revenue Service (IRS) data. Using weighting procedures, the ACS data are controlled to the population (by age, sex, race, Hispanic identity) and housing unit estimates from the Census Bureau's annual population estimate program.

For the 2005 ACS, group quarters were not sampled because of budget restrictions. Thus, the published data contain only the household population. Some data users did not understand these universe differences and made direct comparisons to decennial data that represented the total population.

Although there are a number of considerations for ACS data users, when used properly, the ACS supplies reliable and timely information to help users make better decisions. Many of these issues should be worked out over time as more information is released and data users become more familiar with the data limitations.

Christine Pierce

See also Census; Computer Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Nonresponse; Sampling Error; U.S. Bureau of the Census

Further Readings

Alexander, C. H (2001, October). Still rolling: Leslie Kish's "Rolling Samples" and the American Community Survey. *Achieving data quality in a statistical agency: A methodological perspective.* Proceedings of Statistics Canada Symposium, Hull, Quebec.

American Community Survey Office. (2003). *American Community Survey operations plan.* Washington, DC: U.S. Census Bureau. Retrieved December 8, 2006, from http://www.census.gov/acs/www/Downloads/OpsPlanfinal.pdf

American Community Survey Office. (2005). *Accuracy of the data*. Washington, DC: U.S. Census Bureau. Retrieved December 8, 2006, from http://www.census.gov/acs/www/UseData/Accuracy/Accuracy1.htm

Mather, M., Rivers, K., & Jacobsen, L. A. (2005). *The American Community Survey.* Population Bulletin 60, no. 3. Washington DC: Population Reference Bureau.

U.S. Census Bureau. (2006). *Design and methodology: American Community Survey*. Washington, DC: U.S. Government Printing Office.

American Statistical Association Section on Survey Research Methods (ASA-SRMS)

The Section on Survey Research Methods (SRMS) is a formal section of the American Statistical Association (ASA) that is devoted to encouraging research and the advancement of knowledge in all aspects of survey research. The goals of the SRMS are to promote the improvement of survey practice and the understanding of survey methods in both theoretical and applied research. In 2006, the SRMS was the third-largest section in the ASA, with approximately 1,300 members. All sections of the ASA require that their members first join the ASA.

The SRMS has a relatively short history. In 1974, a group of members of the ASA recognized a need to coordinate and facilitate the study of survey research distinct from other statistical activities. To accomplish this goal, they formed a subsection within the existing Social Statistics Section of the ASA specifically for this purpose. The subsection evolved quickly. It petitioned the ASA to become a full section in 1976, and the petition was approved in 1977 by a vote of the ASA membership. The SRMS began operation as a full section of the ASA in January 1978. In 1990, Irene Hess describes these events and the researchers who helped create the SRMS in an article in *The American Statistician.*

Since its inception as a subsection, the SRMS has identified and fostered research in some areas of special interest to its members. These areas include (a) foundations of sampling; (b) design and execution of sample surveys; (c) nonsampling errors; (d) data collection methods; (e) questionnaire design, evaluation, and testing; (f) analysis and presentation of survey data; (g) education of the public and students on the importance of scientific survey research; (h) publication and dissemination of survey research findings; (i) ethics related to the conduct of survey research; (j) appropriate methods of dealing with respondents and potential respondents; and (k) standards for survey practice.

Disseminating information on survey methods is one of the main functions of the SRMS. The SRMS has been active in a number of ways to disseminate information on survey research methods to a wide audience within the ASA, in the scientific community, and among the public. One approach has been to stimulate the preparation of articles and reports dealing with survey methodology under its auspices. Another approach has been to foster liaisons with persons and organizations publishing papers and monographs on topics of interest in survey methodology. A third approach has been to sponsor topic-oriented workshops, short courses, and conferences of interest to survey researchers.

One of the first such efforts was undertaken in 1976 when the SRMS was still a subsection. A brochure called *What Is a Survey?* was developed and quickly became a key piece of the dissemination effort. The brochure was published several times and was translated into several languages. The brochure was later developed into a series covering specific topics and is still widely used. It is currently available on the SRMS Web site.

The SRMS has also been very active in sponsoring international conferences on specific survey research methods. The first international conference that led directly to an edited monograph was the International Symposium on Survey Methods, cosponsored by ASA Ottawa Chapter, Statistics Canada, and Carleton University in 1980. In 1986, the international conferences sponsored by the SRMS became a continuing series. An international conference has been held every 2 years or so, and nearly all of these conferences resulted in edited monographs of the invited papers. The topics of the conferences have included Panel Samples, Telephone Sampling, Survey Measurement and Process Quality, Business Surveys, Computer Assisted Data Collection, Nonresponse, and Methods for Testing and Evaluating Survey Questionnaires. Nearly all of these conferences were cosponsored by the American Association of Public Opinion Research and the International Association of Survey Statisticians.

At many of the international conferences and the annual Joint Statistical meetings, short courses and tutorials are sponsored by the SRMS. The short courses are presented by survey researchers who are experts in the field and many have recently published books. Topics of the short courses have covered a wide range of methods issues, from questionnaire design to variance estimation with complex samples.

In a more recent and highly effective dissemination effort, the SRMS has scanned all the papers that were prepared for the *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Access to all *Proceedings* papers published by the SRMS going back to 1978 can be obtained without charge from the SRMS Web site. This has been found to be a great benefit to the SRMS members and the survey research community at large.

The SRMS also established and distributes a newsletter for its members. The newsletter provides a forum for keeping SRMS members aware of the activities and concerns of the section as well as informing members of upcoming events, training opportunities, and awards.

Another approach that the SRMS has used to promote interest in survey methods is to award scholarships to students and to honor those who have made important contributions to survey research. For example, the SRMS offers Student Travel Awards to several doctoral students to support their attendance at the ASA annual meeting and attendance at an SRMS short course. In conjunction with other sections of the ASA, the SRMS annually has a competition open to students and postgraduates in survey methodology and related fields, and the winners are given awards to support their attendance at the ASA annual meeting.

Pat Dean Brick

See also American Association for Public Opinion Research (AAPOR)

Further Readings

Hess, I. (1990). History of the Section on Survey Research Methods. *The American Statistician*, *44*(2), 98–100.

Section on Survey Research Methods: http://www.amstat.org/sections/SRMS/index.html

Section on Survey Research Methods. (n.d.). *What Is a Survey*? Retrieved March 4, 2008, from http://www.whatisasurvey.info

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a statistical technique that is used to compare groups on possible differences in the average (mean) of a quantitative (interval or ratio, continuous) measure. Variables that allocate respondents to different groups are called *factors;* an ANOVA can involve one factor (a one-way design) or multiple factors (a multi-way or factorial design). The term *analysis of variance* refers to the partitioning of the total variation in the outcome variable into parts explained by the factor(s)—related to differences between groups, so-called explained or between variation—and a part that remains after taking the factor(s) into account, the so-called unexplained, residual, or within variation.

Consider a one-factor example in which the target population contains respondents from four different ethnic backgrounds (e.g., Chinese, Japanese, Korean, Vietnamese) and the research question is whether these ethnic groups have different average incomes. The null and alternative hypotheses for this example tested with the ANOVA are H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ and H_A: not all μ_j equal, where μ_j $(j = 1, \ldots, 4)$ denote the population mean incomes for the ethnic groups. The test statistic, denoted by F and following an F-distribution, is based on the ratio of the between variation (the variation between the sample group means) and the residual (within groups) variation. A statistically significant result is obtained if the former is large compared to the latter. The conclusion that can be drawn from a significant result is that the mean incomes for the ethnic groups are not all four equal. Of note, no causal conclusions can be made, since this is a nonexperimental study.

In a factorial design, for instance, by the inclusion of gender as a second factor in the previous example hypotheses about main and interaction effects can be tested. A significant main effect of gender implies that the marginal mean incomes of men and women (irrespective of the four ethnic groups) differ. A significant interaction effect of gender and ethnicity on income implies that the differences in mean income between men and women are different among the four ethnic groups.

Some important assumptions underlying the ANOVA are independence of observations and approximately normally distributed residuals, as well as approximately equal residual variances in the subgroups.

Note that the practical conclusions that can be drawn from an ANOVA are somewhat limited. The null hypothesis "all means are equal" is evaluated against the rather uninformative alternative hypothesis stating nothing more than "not all means are equal." Rejecting the null hypothesis in an ANOVA does not inform the researcher about which pairs of means differ from each other. Therefore, an ANOVA is often followed by pair-wise comparisons to further investigate where group differences are found. Since several tests are performed in such a case, the alpha level used per comparison is usually corrected to protect for an increased Type I error probability (post-hoc corrections). Several correction methods are developed, but unfortunately it is not always clear which method should be preferred. Another approach for further investigation of differences between specific means or investigation of a specific structure in the group means is contrast testing.

A second limitation of ANOVA is that directional testing is not possible. An exception is when the ANOVA is applied to a two-mean hypothesis; the ANOVA is then equivalent to the independent samples *t* test. However, it is regularly seen that researchers have specific expectations or theories in terms of the order of the population means. For instance, in a four-group ANOVA the actual hypothesis the researcher is interested in may be: $\mu_1 < \mu_2 < \mu_3 < \mu_4$.

Irene Klugkist

See also Alpha, Significance Level of Test; Factorial Design; *F*-Test; Interval Measure; Level of Measurement; Mean; Null Hypothesis; *p*-Value; Ratio Measure; Significance Level; Subgroup Analysis; *t*-Test; Type I Error; Variance

Further Readings

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

ANONYMITY

Anonymity is defined somewhat differently in survey research than in its more general use. According to the *American Heritage Dictionary*, anonymity is the quality or state of being unknown or unacknowledged. However, in survey research, the concept is more complex and open to interpretation by the various organizations that conduct surveys.

In the form closest to the standard definition, *anonymity* refers to data collected from respondents who are completely unknown to anyone associated with the survey. That is, only the respondent knows that he or she participated in the survey, and the survey researcher can not identify the participants. More often, anonymity refers to data collected in surveys in which the respondents are de-identified and all possible identifying characteristics are separated from the publicly available data. Many survey research organizations provide data and data summaries to individuals outside their organizations. These data are considered anonymous if those outside the survey organization cannot identify the survey participants.

However, for many surveys defined as anonymous, the survey organization could, if needed, identify the respondents. For example, in a survey that uses pure random-digit dial procedures, limited information about the respondent is available to the survey organization. Through the use of various databases, the organization could possibly determine the household associated with the telephone number. Survey organizations would rarely do that.

Survey researchers have developed a number of procedures for designing anonymous surveys. For example, many surveys conducted in classrooms or other gathered events use unnumbered questionnaires and do not contain questions that could identify respondents. For some classroom surveys, identifying information is collected on a sheet separate from the questionnaire.

A procedure sometimes used in postal surveys is to include a return postcard along with return envelope. The unnumbered questionnaire is returned in the envelope, and the postcard is sent separately to let the researchers know that the questionnaire has been returned.

Survey researchers have developed many techniques for conducting completely anonymous surveys. For example, Internet surveys offer multiple methods for anonymous participation. Some surveys may not require authentication to access the survey. Invitations are sent to potential participants but with no control over who participates nor how often. A more sophisticated recruitment method is to completely separate the database used for authentication from the database that contains the survey responses. Another method is for one organization to send the recruiting requests and a second to collect the data.

A similar method can be used for telephone surveys. The telephone numbers can be stored in a database that has no direct link to the survey responses. This method can be used with random-digit dial telephone number samples to further separate the identifying information from the survey responses.

However, the procedures for ensuring anonymity can conflict with other important survey quality control procedures. For example, sending unnumbered paper questionnaires with postcards in postal surveys allows respondents to return the questionnaires but not the postcard. As a result, follow-up requests cannot be limited to nonrespondents only. Respondents who did not return the postcards may believe their first questionnaire did not reach the survey organization and respond a second time.

A similar problem that leads to inappropriate follow-up requests occurs with Internet surveys that do not use authentication. These surveys are open to anyone with Internet access. While some limitations can be applied to prevent unauthorized access, they are minimally effective. The survey data and results are harmed if those not selected for the sample are included in the survey data or respondents participate more than once.

Many survey organizations conduct random checks on survey interviewers to determine whether the interview was conducted and/or was conducted correctly. Survey procedures that ensure anonymity simultaneously prevent these important procedures for verification and monitoring survey quality.

Anonymity is important for the success of surveys under certain conditions. Anonymity can help to protect privacy so that respondents can reveal information that cannot be identified to them. When the survey poses exceptional risks for participants, anonymity may improve cooperation. When a survey asks especially sensitive questions, anonymity will likely improve reporting of stigmatizing behaviors or unpopular attitudes and opinions. Surveys of sexual behaviors, illegal drug use, excessive alcohol use, illegal activities such as tax evasion, and other possibly stigmatizing activities can benefit from providing anonymity to the respondents.

Some participants would be reluctant to discuss attitudes and opinions on such topics as race, politics, and religion unless they believed their responses could not be identified to them. Similarly, respondents have a reduced impetus to provide socially desirable responses in anonymous surveys. For example, respondents may be more willing to admit to negative attitudes toward minority groups if the survey is anonymous.

For these surveys, the risk of exposure or harm to respondents needs to be balanced against the loss of quality control procedures needed to ensure survey integrity. Little empirical evidence is available to indicate the overall importance of anonymity to survey cooperation and survey quality, but survey researchers regularly attempt to use procedures that can ensure anonymity in data collection.

John Kennedy

See also Confidentiality; Ethical Principles; Verification

Answering Machine Messages

Telephone answering machines are devices that automatically answer telephone calls and record messages left by callers when the party called is unable to answer. Within households such devices are often used as "virtual secretaries" to screen unwanted calls or to facilitate communication while away from home. The first automated answering machines became available in the late 1930s in Europe, and the first commercial answering machine was sold in the United States in 1960. It was not, however, until the advent of digital technology in the early 1980s that ownership of telephone answering machines became widespread. Ownership in the United States has increased significantly since then, with more than 70% of households owning a telephone answering machine in 2006. Compared with people who do not have answering machines, owners of these devices typically have higher levels of education and incomes and are more likely to live in households of two or more adults.

Increased ownership of telephone answering machines and their use to screen calls pose a threat to the representativeness of samples in telephone surveys, particularly those based on random-digit dialed designs. More than half of the people who own answering machines say that they or someone else in their household uses the device to screen incoming telephone calls on at least an occasional basis. Households that screen calls are likely to have high family incomes, to be located in suburban or urban areas, and to include young adults with high levels of education. Yet, despite the increased use of answering machines for call screening, many researchers found that households with answering machines can be reached by telephone for survey calls, albeit often after multiple attempts. Fewer than 5% of households appear to screen all of their telephone calls with an answering machine, and when reached, answering machine owners tend to be just as willing to complete surveys as are those without answering machines. Contact with households with answering machines tends to be most successful when calls are made on Saturdays before noon, on Sundays, or on weekdays after 6:00 p.m.

People are not uniform, however, in how they use telephone answering machines. People with on-the-go lifestyles tend to use telephone answering machines to stay in contact and facilitate communication. This finding led some researchers to hypothesize that scripted messages left on such devices may prepare the household for a later call or even encourage a prospective respondent to return the call free of charge to complete the interview. If successful, such an approach would help to reduce the level of nonresponse in telephone surveys.

However, empirical research on the effectiveness of leaving messages on answering machines to improve survey participation is mixed. For surveys that involve a list of sample members whose names are known, leaving messages can be effective at improving survey participation. Such messages appear to work best if the message is tailored to include the sample member's name. Several random-digit dialed telephone surveys conducted in the early 1990s also showed that leaving messages on telephone answering machines could significantly improve response rates by 3 to 4 percentage points. However, more recent studies conducted at the state and national levels using random-digit dialed sample designs found no difference in the contact or completion rates of households that were left a message and those that were not. The strategy does not appear effective for two reasons. First, the percentage of households with which this technique can be used is limited, since messages can be left only at households with answering machines that are set to receive messages. Although telephone answering machines are in more than 70% of households, not all of these machines are ready to receive messages every time a survey call is made. Second, only a small percentage of respondents within households hear the message and are positively influenced to participate in the survey. It may be that people in households with multiple adults or teenagers sort through and listen to telephone messages in much the same way they sort through mail: one person tends to sort and screen for the rest of the household. It is likely that one person (perhaps simply the first person home each day) will listen to all of the telephone messages and relay to others in the household what is deemed to be important information. Unsolicited calls from researchers are probably not at the top of that priority list. As a result, with the exception of the person who sorts the messages, probably few other adults in the household hear them.

In addition, leaving messages on telephone answering machines has real costs. Leaving messages takes interviewer time, both to listen to the greeting on the answering machine and message and to leave the notice about the survey. This added time increases costs and does not appear to produce positive returns in the form of either lower nonresponse rates or less interviewer labor.

Michael W. Link

Further Readings

Link, M., & Mokdad, A. (2005). Leaving answering machine messages: Do they increase response rates for the Behavioral Risk Factor Surveillance System? *International Journal of Public Opinion Research*, *17*, 239–250.

Link, M. W., & Oldendick, R. W. (1999). Call screening: Is it really a problem for survey research? *Public Opinion Quarterly*, *63*, 575–589.

Oldendick, R. W., & Link, M. W. (1994). The answering machine generation. *Public Opinion Quarterly*, *58*, 264–273.

Tuckel, P., & Feinberg, B. (1991). The answering machine poses many questions for telephone survey researchers. *Public Opinion Quarterly*, *55*, 200–217.

Approval Ratings

Approval ratings are a particularly versatile class of survey questions that measure public evaluations of a politician, institution, policy, or public figure as well as judgments on public issues. This type of question was first developed by the Gallup Organization in the late 1930s to measure public support for the U.S. president. Today, the presidential job approval question is believed to be the single most frequently asked question in political surveys. Many members of the political community, journalists, and academics consider the job approval question to be among the most reliable and useful barometer of a president's public standing.

Basic Question Format

While versions of the job approval question were asked by George Gallup in the late 1930s, the modern form of the presidential approval question was finally adopted by Gallup in the mid-1940s, according to the Gallup Organization. Since then, the Gallup wording remains unchanged, giving journalists and academics an historic record of public evaluations of their presidents for more than 60 years.

The basic form reads: *Do you approve or disapprove of the way (*name of president*) is handling his job as president?* Some polling organizations use slightly different wording, but most have adopted the Gallup language, in part so they can compare the results with Gallup's historic data without having to worry about the effect of wording differences. A variation of the question is frequently used to measure a president's performance in specific domains, as with this trend question asked by *The Los Angeles Times*: *Do you approve or disapprove of the way George W. Bush is handling the war on terrorism?*

The question's basic format is easily altered to evaluate the performance of other public officials or institutions, such as Congress, individual members of a president's cabinet, or state and local officials, as well as other prominent leaders. It also is a useful measure of public attitudes toward government programs or policies and frequently is used to measure attitudes toward a range of nonpolitical issues, such as this question by *USA Today* and Gallup: *Do you approve or disapprove of marriage between blacks and whites?*

Polling organizations often include language that measures the intensity of approval or disapproval, as with this approval question asked in 2005 by the Pew Center for the People and the Press: *There is now a new Medicare law that includes some coverage of prescription drug costs. Overall, would you say you strongly approve, approve, disapprove, or strongly disapprove of the way Medicare will now cover prescription drug costs?* These strength-of-support measures allow survey respondents to indicate a degree of approval or disapproval, and thus are more sensitive to change in public attitudes. For example, declining public support for elected officials is often first seen as a decline among those who strongly approve of him or her and a comparable increase in those who somewhat support the official, with little or no decline in the overall support.

Presidential Approval Ratings

President George W. Bush has the distinction of having the highest as well as one of the lowest overall job approval ratings in Gallup polls of any president in the modern era. In an ABC survey conducted 4 weeks after the terrorist attacks of September 11, 2001, Bush recorded a 92% job approval rating, the highest job performance rating ever achieved by an American president in a major national poll. Other polling organizations also recorded historic highs for Bush in this time period. Coincidentally, Bush's father, George H. W. Bush, achieved the second-highest job approval rating in Gallup surveys, 89%, in February 1991, after the quick Allied victory in the Gulf War. Both numbers stand as striking illustrations of the power of the presidential job rating to measure *rally effects* in American politics, that is, the tendency of the public to rally behind their leader in times of national crisis. In a survey conducted by *The Washington Post* and ABC News the week before the 9/11 terrorist attacks, George W. Bush's job approval rating stood at 55%, 35 percentage points below his approval rating in a *Post*/ABC survey 2 weeks after the attacks.

As these numbers suggest, times of war and national crisis have produced sharp spikes in presidential approval. Other presidents with high job approval ratings in Gallup polls include Franklin Delano Roosevelt, who had an 84% approval rating in January 1942, after the Japanese attacked Pearl Harbor and Germany declared war on the United States. Harry S Truman had an overall job approval rating of 87% in June 1945, after the end of World War II in Europe and just before Japan surrendered. (The Gallup question, however, was slightly different in that it asked whether people approved or disapproved of the way Roosevelt *is handling his job as President today*. The word *today* was dropped three years later.)

Truman also has the distinction of being the president with the lowest job approval rating ever recorded by Gallup: 22% in February 1952, a consequence of public dissatisfaction with the Korean War. At the climax of the Watergate scandal in the summer of 1974, Richard Nixon's approval rating was 24%, while George W. Bush matched Nixon's low in a Reuters-Zogby survey in October 2007. Scandal does not automatically send a president's job approval rating plummeting. Most political observers expected that President Bill Clinton's job approval rating would collapse after details of his affair with White House intern Monica Lewinsky were revealed. In fact, his approval rating dropped insignificantly, if at all, in most public polls and quickly rebounded; whatever his failings as a person, the public continued to give Clinton high marks for his on-the-job performance as president.

Retrospective Judgments

Approval questions sometimes are used to measure the public's retrospective judgments. *USA Today* and Gallup asked this question in 1995 on the 50th anniversary of the end of World War II: *As you may know, the United States dropped atomic bombs on Hiroshima and Nagasaki in August 1945 near the end of World War II. Looking back, would you say you approve or disapprove of using the atomic bomb on Japanese cities in 1945?* Such a format has provided an interesting view of the American public's retrospective judgment of its presidents. When Gallup asked the public in 2002 if they approved or disapproved of the job done by each of the presidents in the post–World War II era, President John F. Kennedy topped the list with 83% approval, followed by Ronald Reagan (73%), and Jimmy Carter (60%).

The retrospective approval question is regularly asked by Gallup. The results over time suggest that an elected official's job approval rating can change significantly even after he or she leaves office. In 2002, Gallup found that 69% of the public approved, in retrospect, of the job that George H. W. Bush had done as president. But in 2006, the elder Bush's job rating had declined from 69%, third-highest behind Kennedy and Reagan, to 56%. Conversely, President Clinton's retrospective job approval rating increased from 51% in 2002 to 61% four years later.

Question Order Effects

Pollsters have found that job approval questions can be particularly sensitive to question order effects. For example, the overall job approval rating of Congress can be significantly different if the question is asked in a survey before or after a series of questions that ask people to evaluate how effective lawmakers were in dealing with a set of controversial issues. Presidential approval ratings tend to be higher when the question is asked first in a survey compared to when they are asked later in the survey after various policy issues and evaluations. That is why the presidential job approval rating and other approval questions typically are asked near or at the beginning of a survey.

Richard Morin

See also Likert Scale; Question Order Effects

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

AREA FRAME

An area frame is a collection of well-defined land units that is used to draw survey samples. Common land units composing an area frame include states, provinces, counties, zip code areas, or blocks. An area frame could be a list, map, aerial photograph, satellite image, or any other collection of land units. Area frames play an important part in area probability samples, multi-stage samples, cluster samples, and multiple frame samples. They are often used when a list of ultimate sampling units does not exist, other frames have coverage problems, a geographically clustered sample is desired, or a geographic area is the ultimate sampling unit.

Plot and Grid Area Frames

There are two types of area frames: grid frames and plot frames. The distinction between a grid and plot frame is based on the analytical goal of the survey rather than the structure of the frame. *Plot frames* contain ultimate sampling units that are observed in their entirety, whereas *grid frames* contain land units that will be further divided and sampled at further stages.

Plot frames are often used in agricultural and environmental surveys in which measurements are taken on a piece of land. For example, consider a survey designed to estimate pollutants in a stream. After obtaining a map of the stream, one could partition the stream into 3-foot-by-3-foot square plots. If a sample of plots is selected and the pollutants in each sample plot are measured, then the map of 3-foot-by-3-foot square plots is a plot frame, because the entire plot is enumerated.

Sometimes is it desirable to select a sample of units within geographic areas. In grid frames, geographic clusters of sample units compose the frame. The geographic clusters are first sampled. Then a sample is selected from units within the sampled clusters.

Use of Area Frame in Multi-Stage Sampling

Grid area frames play a central role in multi-stage sampling. At every stage of selection except the final stage, a different area frame is used. For example, consider a survey designed to estimate the median income of all households in a city. In the United States, one possible area frame for the first stage of sample is a list of all census tracts. After selecting a set of tracts, one could construct a second area frame of all census blocks within the selected tracts. Blocks that are not in selected counties are not considered a part of the sampling frame because they do not have a chance of selection.

Before selecting the final stage of households in sample blocks, a list of households within the blocks needs to be built. Field staff often perform this role by listing all households within the selected blocks; although the list of addresses could be obtained from an administrative list. In the final stage of sampling, the list of housing units is an example of a list frame rather than an area frame. However, sometimes geographically clustered lists built from a field enumeration are referred to as an area frame.

Reasons to Use Area Frames

When a satisfactory list frame is not available, an area frame may be the best alternative. For example, consider a survey of homeless adults in a large city. In the absence of a list of homeless people in the city, one could construct an area frame of city blocks that would cover the entire population. In such a case one might also want to use a second frame of people staying in a homeless shelter to supplement the area frame.

Sometimes area frames are used to enhance an imperfect frame. For example, a national survey of households might use a frame of telephone numbers supplemented by an area frame. The sample drawn from the telephone list will not cover households without telephone service. However, constructing the entire survey from an area frame may be too expensive. Thus some surveys use an area frame to enhance a frame with known coverage deficiencies.

For surveys involving personal interviews, geographic clustering provides a way to reduce field costs. For example, it is more efficient to interview four different households in the same city block than four different households spread out in a large area. Selecting a multi-stage sample from area frames is the most common way to obtain a geographically clustered sample.

Finally, plot area frames are used when the geographic area is of interest. For example, area frames are widely used in measuring the coverage of address lists. To do so, a sample of geographic areas is selected from a plot area frame. Then, field staff lists all the addresses in the sample areas, which are then compared to the list frame to measure coverage.

Area Frame Construction

In many cases it is possible to enhance an area frame with a wealth of auxiliary data that can be used in stratification, allocation, and sampling. Accurate estimates of the estimated measure of each geographic unit's size is of particular importance in the case of area probability sampling.

Area frames should cover the entire population and partition it into mutually exclusive geographic units. Indeed, the best frames have well-defined boundaries because poorly defined boundaries are likely to lead to coverage problems. For surveys that make estimates based on political boundaries such as counties or cities, some tradeoff usually has to be made between visible geographic boundaries and "invisible" political boundaries.

Besides being clearly defined with visible boundaries, area frames should be up-to-date and accurate. Changes in the political geography such as city annexations as well as changes in the physical geography such as changing rivers, tree rows, and roads should be reflected in the area frame boundaries. Out-of-date boundaries can cause confusion in the field, increasing cost, coverage bias, and coverage variance.

Last, each unit in the area frame should be unique. For example, an area frame of counties must also include the state name, otherwise there would be no way of differentiating Montgomery County, Alabama, from Montgomery County, Maryland.

Timothy Kennel

See also Area Probability Sample; Auxiliary Variable; Cluster Sample; Coverage; Multiple-Frame Sampling

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley.

Jessen, R. J. (1978). *Statistical survey techniques.* New York: Wiley.
Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys.* New York: Wiley.
Yeates, F. (1981). *Sampling methods for censuses and surveys.* New York: Macmillan.

AREA PROBABILITY SAMPLE

An area probability sample is one in which geographic areas are sampled with known probability. While an area probability sample design could conceivably provide for selecting areas that are themselves the units being studied, in survey research an area probability sample is usually one in which areas are selected as part of a clustered or multi-stage design. In such designs, households, individuals, businesses, or other organizations are studied, and they are sampled within the geographical areas selected for the sample. An example of a survey that uses area probability sampling in the United States is the Current Population Survey (CPS).

Terminology

There are several terms that are used in relation to area probability sampling that are not frequently used except in area probability and other multi-stage sampling designs. In area probability samples, the units formed for selection at the first stage are called *primary sampling units* (PSUs) and those for the second stage of selection are called *secondary sampling units* (SSUs). The units that are actually selected at these stages are called, respectively, *primary* and *secondary selections*. If there are more than three stages, the units for the third stage may be called *tertiary selection units* or *third-stage selection units*. The final unit to be selected is called the *ultimate sampling unit*.

PSUs, SSUs, and perhaps other units are often selected using probability proportional to size (PPS) methods. In these cases, each selection unit is assigned a measure of size (MOS). The MOS usually represents the size of the study population found in the unit. The MOS may be known or estimated or may be a function such as the square root of the population total or a composite (e.g., the sum of the total number of males plus 1.5 times the total number of females).

Reasons for Using Area Probability Designs

Many considerations can affect the choice of an area probability design for a study. One reason to use this approach could be that there is no available satisfactory list of the study population that can serve as a sampling frame. In other cases, the researchers may desire to use data about the areas as correlates in analysis of other data collected from persons or establishments. Often the choice is driven by the fact that the data being collected are best obtained (or can only be obtained) through personal contact with, or observation of, members of the population being studied. For example, (a) questionnaire items may require that the respondent be presented with visual cues as can be done in face-to-face interviewing; (b) the study requires that medical specimens be taken or anthropometric measurements be made; (c) the data collection involves observing behaviors, situations, or the physical environment.

If personal contact is required, cost considerations may make a clustered or multi-stage area probability sample design the most efficient, if not the only feasible design. For instance, if the survey is to collect data through personal contact with 3,000 adults in the United States, a simple random sample (or other unclustered design), even if possible, would be prohibitively expensive. An example of a more affordable design would be collecting data on 30 adults in each of 100 relatively compact areas such as metropolitan areas, counties, cities, towns, or similar administrative areas.

Disadvantages of Area Probability Samples

There are two major disadvantages to using an area probability sample: (1) the increase in variance, often called a design effect (*deff*) that comes from the use of multi-stage or clustered designs, and (2) the increased cost that is mostly associated with using in-person data collection (although not all studies with area probability sample designs use in-person data collection).

The design effect due to clustering arises from the fact that the units of observation in the study, be they individuals, households, or businesses, are not selected independently, but rather their selection is conditional on the cluster (in this case a geographic area) in which they are found being selected. In area probability sampling, the design effect of clustering can be small for some variables (estimates of gender and age, and some attitudinal measures), moderate for others (economic variables), and substantial for others (estimates of the prevalence of racial or ethnic groups).

The increased cost can come from having to have interviewers visit homes or businesses, but it can also come from the sampling process itself if part of the sampling frame must be developed by having field workers travel to selected areas and compile lists of addresses.

Procedures for Designing and Selecting Area Probability Samples

The first step in designing an area probability is defining the study population in geographic terms (e.g., adults living in the United States; students attending charter schools in the state of New York; or registered voters in the Mexican state of Zacatecas). The second step is to find or develop a sampling frame or frames, since the process often involves finding or developing a frame for each stage of selection. The frames should comprise lists of the sampling units at each stage, with all the information needed to stratify and implement the selection plan. The initial list may not correspond exactly to the sampling units that will be defined, but it should contain the information needed to create the frame once the sampling units are defined. For example, a list of counties or cities could be used to compile a frame of PSUs, some of which would include multiple counties or cities.

Since the size of the sampling units is important for selecting the sample in most area probability designs, data about the size of each PSU should be available. In addition, geography and economic and demographic measures may be needed. In most countries there will be lists available from government agencies that will serve as a frame for the PSUs. Constructing frames for the subsequent stages of selection may require more work, and depending on study needs, will call for creativity.

The next several steps involve defining sampling units and the strata within which they are to be sampled. What geographic areas will comprise the PSUs, SSUs, and other sampling units? Attention should be paid to the size of the units. As a rule of thumb, an area probability sample should have a minimum of 30 to 50 PSUs; a hundred or more are preferred for large studies. If the PSUs are too large, the sample may not be able to include a desirable number of selections. On the other hand, small PSUs may be more homogeneous than desired. A good approach is to have PSUs large enough that sampling the SSUs and subsequent units can introduce heterogeneity into the sample within each PSU.

After defining the PSUs, at least in general terms, strata are defined. Part of the stratification process involves defining "certainty selections," that is, PSUs that are large enough that they are certain to be selected. Each certainty PSU becomes its own stratum. One can think of certainty selections in terms of a sampling interval for systematic selection. To this end, define the interval (I) as the sum of the MOS for all PSUs in the population (MOSTOT) divided by the number of PSUs to be selected (n_PSU):

$$I = MOSTOT/n\_\text{PSU}.$$

Thus, any PSU with an MOS at least as large as I would be certain to be selected. If there are certainty selections, then it is advisable to set the cutoff for designating a PSU as a certainty selection as a fraction of I (perhaps 0.8 times I). The reason for this is that once the certainty PSUs are removed from the population, the sum of the MOS becomes smaller, and possibly additional PSUs will become large enough to be certainty selections: the sum of the remaining MOS can be designated MOSTOT* and the number of PSUs to be selected after the certainty selections are made as n_PSU_noncert. If one calculates a new sampling interval $I^* = \text{MOSTOT}^*/n\_\text{PSU\_noncert}$, it is possible that there will be new certainty selections the MOS for which is equal to or greater than I^*. Setting the certainty cutoff as a fraction of I usually avoids the problem of having to go through several iterations of removing certainty PSUs from the pool.

Once all certainty selections have been defined, the other PSUs on the frame are grouped into strata. As for any study, the strata should be related to study objectives, especially if subgroups of the population are to be oversampled. Area probability samples are often stratified geographically. The number of strata for the first stage is limited by the number of primary selections to be made. To estimate sampling variance, each stratum should be allocated at least two primary selections. Some deeply stratified designs call for one selection per stratum, but in such a design, strata will have to be combined for variance estimation.

The process just described for PSUs is then repeated for SSUs, third-stage units, and so on. It is only necessary to define SSUs within PSUs that are actually selected for the sample. SSUs within certainty PSUs are treated as primary selections for estimating sampling error (the certainty PSUs are treated as strata). The selection of units within PSUs depends on the purposes of the study. Oversampling may be accomplished through the use of stratification or giving extra weight when creating the MOS to the group(s) to be oversampled. If no oversampling is desired, it is possible, by using PPS at all stages, to have nearly equal probabilities of selection for the ultimate sampling units.

The sampling frames at the final or next-to-final stages often require substantial field labor. For example, field workers may have to visit the sampled areas and make lists, based on visual inspection, of dwelling units or businesses. In addition to taking the cost of listing into account, area probability sample designs must be flexible in case MOS at the later stages are substantially incorrect—whole blocks may have been destroyed by natural disasters or to make way for new construction, or the new construction may have taken place and the area contains many more dwellings or businesses than were anticipated. If an area has grown substantially, it may have to be subdivided before listing—essentially adding another stage of sampling.

Hypothetical Example of an Area Probability Design

In the United States, many large ongoing surveys operated or funded by the federal government use area probability designs. These include surveys of households or individuals as well as studies of businesses and other establishments. The subject areas of these surveys range from labor force participation to health status to energy consumption and other topics. Rather than try to examine the details of such sample designs, what follows is a hypothetical (generic) example of a sample design for a survey in which adults living in households comprise the target population and in-person data collection is required. Although there could be more stages of sampling, this example deals with four: (1) at the first stage, PSUs will be defined as "large" geographic areas; (2) in the second stage, somewhat smaller geographic areas will be defined as SSUs; (3) the third-stage units will be households identified within the SSUs; and (4) the fourth-stage (in this case ultimate) units will be adults identified within households.

If the survey were conducted in the United States, the PSUs very likely would be defined as metropolitan areas or counties. (Larger units, such as states, would probably be inefficient for most surveys.) The sampling frame, a list of all PSUs, would be stratified, possibly using a combination of variables such as region of the country, population density, economic and demographic characteristics. The stratifying variables would depend in part on whether the design was a general purpose one (to be used for many, perhaps unrelated studies) or a more specific one (such as for a study of a particular ethnic group).

SSUs in the United States might comprise areas defined for the U.S. Decennial Census, such as tracts, block groups, or blocks. The sampling frame for the SSUs would probably be electronic or other lists of these units obtained from the U.S. Census Bureau. The frame of SSUs should be stratified within each PSU; often the stratifying variables are similar to those used in sampling PSUs.

To create sampling frames of households within the SSUs, lists of dwellings or addresses are compiled, possibly by having field workers record the addresses on forms or enter them on portable computers. It is also possible to define sets of addresses based on postal delivery files or other administrative lists. These lists (whether created by study staff or obtained from postal or other administrative records) may be incomplete; thus, procedures need to be devised so that dwellings not on the list have a chance of being selected. One such method is the *half-open interval method*, in which unlisted units within a certain interval are given a known chance of selection.

The list of addresses or dwellings comprises the sampling frame for selecting households. However, at this point the study usually introduces two-phase sampling, since the list must be screened to determine if the dwellings identified on the list contain eligible households. This screening might be done on all units listed or on a subsample. For this example, we will assume that all listed units are screened. Examples of addresses that would not be eligible for this hypothetical survey include apparent dwellings that are actually businesses; vacant or uninhabitable structures; dwellings for which the group of people living there do not meet the definition of a household (for example a halfway house for recovering alcoholics or inmates close to being released from prison); or dwellings that do not contain an adult.

For this hypothetical example, the study will attempt to conduct interviews at all dwellings that contain households with adults; this is a likely scenario since it can reduce nonresponse if the interview is attempted at the same time as the household is screened. At this point, the design might call for attempting to interview (or otherwise collect data about) all adults in the household or for random selection of one adult to be interviewed.

John Hall

See also Cluster Sample; Current Population Survey (CPS); Cutoff Sampling; Design Effects (*deff*); Face-to-Face Interviewing; Field Work; Half-Open Interval; Multi-Stage Sample; Probability of Selection; Probability Proportional to Size (PPS) Sampling; Sampling Frame; Sampling Variance; Simple Random Sample; Strata; Stratified Sampling; Target Population; Unit; Variance Estimation

Further Readings

Aquilino, W. S., & Wright, D. L. (1996). Substance use estimates from RDD and area probability samples: Impact of differential screening methods and unit nonresponse. *Public Opinion Quarterly*, *60*(4), 563–573.

Haner, C. F., & Meier, N. C. (1951). The adaptability of area-probability sampling to public opinion measurement. *Public Opinion Quarterly*, *15*(2), 335–352.

Henry, G. (1990). *Practical sampling.* Newbury Park, CA: Sage.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons.

Nielsen Media Research. (2008). *Sampling the population.* Retrieved March 3, 2008, from http://www.nielsenmedia.com/nc/portal/site/Public/menuitem.55dc65b4a7d5adff3f65936147a062a0/?vgnextoid=bc0e47f8b5264010VgnVCM100000880a260aRCRD

Attenuation

Attenuation is a statistical concept that refers to underestimating the correlation between two different measures because of measurement error. Because no test or other measurement of any construct has perfect reliability, the validity of the scores between predictor and criterion will decrease. Hence, when correlating scores from two survey instruments, the obtained correlation may be substantively lower if the score reliabilities from both instruments are suspect. Therefore, Charles Spearman proposed the following "correction for attenuation" formula, estimating the correlation between two measures if the scores on both had perfect reliability:

$$r_{xyc} = \frac{r_{xy}}{\sqrt{r_{xx} * r_{yy}}}.$$

In this formula, r_{xyc} is the correlation between the predictor (x) and the criterion (y) corrected for attenuation; r_{xy} is the correlation between the predictor and criterion scores; r_{xx} is the reliability of the predictor scores; and r_{yy} represents the reliability of the criterion scores.

Suppose the correlation between scores on self-esteem and anger scales is .30. If the reliability (e.g., Cronbach's alpha) of the scores from the self-esteem inventory is .80 and the reliability of the scores from the anger inventory is .90, then the correction for attenuation would be equal to the following:

$$.35 = \frac{.30}{\sqrt{.80 * .90}}.$$

Because the reliabilities of the scores from the self-esteem and anger scales are high, there is little correction. However, suppose the score reliabilities for the anger and self-esteem inventories are extremely low (e.g., .40). The correction for attenuation would escalate to .75. If the square root of the product of the reliabilities were less than .30, then the correction for attenuation would be greater than 1.0!

However, rather than correcting for score unreliability in both measures, there are times in which one would correct for score unreliability for either the predictor or criterion variables. For example, suppose the correlation between scores from a job interview (x) and from a personnel test (y) is equal to .25, and assume that the reliability of the personnel test is .70. If one corrected only for the score unreliability of the criterion, then the following equation would be used:

$$r_{xyc} = \frac{r_{xy}}{\sqrt{r_{yy}}}.$$

In this case, the correction for attenuation would equal .30. One could also use a similar equation for correcting the predictor variable. For example, suppose the correlation between scores from a personnel test (x) and the number of interviews completed in a week (y) is equal to .20 and the score reliability of the personnel test is .60. The correction for attenuation would equal .26, using the following equation for correcting only for the score reliability of the predictor variable:

$$r_{xyc} = \frac{r_{xy}}{\sqrt{r_{xx}}}.$$

Paul Muchinsky summarized the recommendations for applying the correction for attenuation. First, the corrected correlations should neither be tested for statistical significance nor should they be compared with uncorrected validity coefficients. Second, the correction for attenuation does not increase predictive validity of test scores. Donald Zimmerman and Richard Williams indicated that the correction for attenuation is useful given high score reliabilities and large sample sizes. Although the correction for attenuation has been used in a variety of situations (e.g., meta-analysis), various statisticians have suggested caution in interpreting its results.

N. Clayton Silver

See also Correlation; Cronbach's Alpha; Reliability; Validity

Further Readings

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63–75.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

ATTITUDE MEASUREMENT

Researchers from a variety of disciplines use survey questionnaires to measure attitudes. For example, political scientists study how people evaluate policy alternatives or political actors. Sociologists study how one's attitudes toward a social group are influenced by one's personal background. Several different methods, including multi-item measures, are used to measure attitudes.

Question Format

People hold attitudes toward particular things, or *attitude objects*. In question format, an attitude object is presented as the stimulus in an attitude question, and respondents are asked to respond to this stimulus. Consider the following question:

> Do you approve, disapprove, or neither approve nor disapprove of the way the president is handling his job?

The attitude object in this question is the president's handling of his job. The respondents must consider what they know about how the president is handling his job and decide whether they approve, disapprove, or neither approve nor disapprove. Another possible closed-ended format is to turn the question into a statement, and ask the respondents whether they agree or disagree with a declarative statement, for example, *The president is doing a good job.* However, some research indicates that the agree–disagree format produces "acquiescence bias" or the tendency to agree with a statement regardless of its content. Yet another closed-ended format is to ask the respondents to place themselves on a continuum on which the endpoints are labeled. For example, one could ask, *How do you feel the president is handling his job?* and ask the respondents to place their opinions on a scale, from 0 being poor to 10 being excellent.

Researchers measuring attitudes must decide how many scale points to use and how to label them. Five to seven scale points are sufficient for most attitude measures. Assigning adjectives to scale points helps define their meaning, and it is best if these adjectives are evenly spaced across the continuum.

Sometimes a researcher wants to understand the preferences of respondents in more depth than a single closed-ended question will allow. One approach for this purpose is to ask the question in an open-ended format such as, *If the Democratic Party were a person, what traits would you use to describe it?* Here, the Democratic Party is the attitude object or stimulus. An advantage of the open format is that the answers are not limited to the researchers' own categories. The answers to such a question will provide insights into whether or not the respondent holds positive, negative, or conflicted attitudes toward the attitude object (the Democratic Party, in this example). However, open-ended responses can be very time consuming to code and analyze. Alternatively, one can list a series of attributes and ask the respondent to rank them. This is easier to analyze but can be cognitively complex if respondents are asked to rank too many items.

Two other important considerations for the response options are whether or not to include a "No opinion" option and/or a middle option. Research suggests that more respondents will use both of these options when they are explicitly offered than when it is left up to respondents to volunteer such responses on their own. Research has also shown that many respondents are willing to offer opinions on obscure or fictitious issues, especially when a "no opinion" option is not offered as an explicit response choice. However, other research suggests that an explicit "no opinion" option may encourage individuals who do have attitudes to not report them. In some measurement contexts, using a middle response choice that conveys a position of noncommitment toward the attitude object makes sense. However, those who have less intense feelings or views about an issue are disproportionately influenced by the inclusion of a middle option. For this reason, the middle option is sometimes omitted, and attitude strength instead is measured with a separate question.

Multi-Item Scales

Another way to measure attitude strength is by using multi-item scales. All scaling procedures require the creation of a pool of items from which a respondent is asked to select a final set according to some criteria. For example, Thurstone scaling first requires a set of judges to rate or compare several statements on a continuum from unfavorable to favorable toward the attitude object. The judges' scores for each statement are then averaged to align the statements along the attitude continuum. These average scores from the judges become the scale values for each statement. Next, the statements are administered to the respondents. The respondents are asked whether they agree with the statements. The respondents' score is then a function of the scale values for the statements that the respondents agreed with.

Guttman scaling is similar, except that it requires an assumption about the pattern of responses that is rarely met in practice. The assumption is that the data set associated with a Guttman scale has a cumulative structure, in the following sense: For *any* two persons in the observed sample, one of them would exhibit all the manifestations of the trait that the other person would, and possibly additional ones. That is, there would be no two persons in the sample with one person higher than the other in one variable but lower than the other in another variable.

Thurstone and Guttman approaches require a significant amount of developmental work. In contrast, Likert scales are much easier to construct. Typically, the researcher selects the statements that correlate the strongest with the sum of the responses to all the statements. The final scale is administered by asking the respondents to respond to the selected statements using a traditional 5- or 7-point agree–disagree response scale. The respondent's attitude is then represented by the sum of the responses to the individual statements or some weighted combination of responses. Although multi-item scales increase the reliability of a measure, thereby reducing measurement error, a disadvantage is that they can seem redundant to some respondents.

Evaluating and Refining Attitude Measures

All attitude questions should be carefully constructed regardless of the format that is used to measure them. The questions should be pretested, using techniques such as cognitive interviewing to ensure that respondents are interpreting the questions as intended. Split-half experiments also can be useful for pretesting alternative versions of a question. It is important to pretest attitude measures in a realistic situation since it is known that attitude questions can be sensitive to the context in which they are asked.

Aaron Maitland

See also Acquiescence Response Bias; Attitudes; Attitude Strength; Coding; Cognitive Interviewing; Context Effect; Guttman Scale; Likert Scale; Nonattitude; Open-Ended Question; Questionnaire Design; Split-Half

Further Readings

Anderson, A. B., Basilevsky, A., & Hum, D. P. (1983). Measurement: Theory and techniques. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 231–287). San Diego, CA: Academic Press.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Newbury Park, CA: Sage.
DeVellis, R. F. (2003). *Scale development.* Thousand Oaks, CA: Sage.
Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

ATTITUDES

Attitudes are general evaluations that people hold regarding a particular entity, such as an object, an issue, or a person. An individual may hold a favorable or positive attitude toward a particular political candidate, for example, and an unfavorable or negative attitude toward another candidate. These attitudes reflect the individual's overall summary evaluations of each candidate.

Attitude measures are commonplace in survey research conducted by political scientists, psychologists, sociologists, economists, marketing scholars, media organizations, political pollsters, and other academic and commercial practitioners. The ubiquity of attitude measures in survey research is perhaps not surprising given that attitudes are often strong predictors of behavior. Knowing a person's attitude toward a particular product, policy, or candidate, therefore, enables one to anticipate whether the person will purchase the product, actively support or oppose the policy, or vote for the candidate.

What Is an Attitude?

An *attitude* is a general, relatively enduring evaluation of an object. Attitudes are *evaluative* in the sense that they reflect the degree of positivity or negativity that a person feels toward an object. An individual's attitude toward ice cream, for example, reflects the extent to which he or she feels positively toward ice cream, with approach tendencies, or negatively toward ice cream, with avoidance tendencies. Attitudes are *general* in that they are overall, global evaluations of an object. That is, a person may recognize various positive and negative aspects of ice cream, but that person's attitude toward ice cream is his or her general assessment of ice cream taken as a whole. Attitudes are *enduring* in that they are stored in memory and they remain at least somewhat stable over time. In this way, attitudes are different from fleeting, momentary evaluative responses to an object. Finally, attitudes are specific to particular objects, unlike diffuse evaluative reactions like moods or general dispositions.

Given this conceptualization, attitudes are most commonly measured by presenting respondents with a bipolar rating scale that covers the full range of potential evaluative responses to an object, ranging from extremely negative to extremely positive, with a midpoint representing neutrality. Respondents are asked to select the scale point that best captures their own overall evaluation of a particular attitude object.

In the National Election Studies, for example, respondents have often been asked to express their attitudes toward various groups using a "feeling thermometer" ranging from 0 (very cold or unfavorable) to 100 (very warm or favorable), with a midpoint of 50 representing neither warmth nor coldness toward a particular group (e.g., women). By selecting a point on this scale, respondents reveal their attitudes toward the group.

How Are Attitudes Formed?

At the most general level, attitudes can be formed in one of three ways. Some attitudes are formed primarily on the basis of our cognitions about an object. For example, we may believe that a particular brand of laundry detergent is reasonably priced, removes tough stains, and is safe for the environment. On the basis of these and other beliefs, we may come to hold a positive attitude toward the detergent. This attitude would be cognitively based.

In contrast, some attitudes are based on few or no cognitions. Instead, these attitudes are based primarily on our affective reactions to an object. Instead of deriving our attitude toward a laundry detergent from our beliefs about its various attributes, for example, we may form an attitude toward it on the basis of the feelings that we associate with the detergent. An advertisement for the detergent that makes us laugh, for example, may leave us feeling positive toward the detergent, even though the advertisement conveyed no substantive information about the detergent.

Attitudes can also be derived from our past behaviors. Sometimes this occurs through self-perception processes. In much the same way that we often infer other people's attitudes from the behaviors they perform, we sometimes look to our own behavior to determine our attitudes. When asked about our attitude toward a particular laundry detergent, for example, we may canvass our memory for relevant information. One thing that we may recall is our past behavior regarding the detergent. We may remember, for example, that we have purchased the detergent in the past. On the basis of this behavior, we may infer that we hold a positive attitude toward the detergent, even if we know nothing else about the product.

In addition to these self-perception processes, there is another way in which our past behavior can influence our attitudes. Instead of inferring our attitudes from our past behavior, we sometimes modify our attitudes to bring them into line with behaviors we have performed. This occurs because, in general, people prefer to exhibit consistency. In fact, according to cognitive dissonance theory, people are very uncomfortable when they recognize an inconsistency among their cognitions, and they are highly motivated to reduce this discomfort. For example, the knowledge that we have performed a behavior that is incongruent with our attitude often produces a state of tension. Resolving this tension requires that we eliminate the inconsistency. Because the behavior has already been performed, it is often easiest to do this by changing the attitude to bring it into line with the behavior. And indeed, a large body of evidence suggests that people often do change their attitudes to make them more consistent with past behaviors.

Why Do People Hold Attitudes?

Attitudes are ubiquitous—we hold them toward people, places, and things, toward concepts and ideas, and toward the vast array of stimuli in our environment. Why do we store these evaluations in memory? Attitudes are believed to serve a number of important psychological functions. Perhaps the most fundamental of these is a "utilitarian" function. Attitudes enable us to efficiently and effectively obtain rewards and avoid punishment by summarizing the positive or negative connotations of an object, guiding our behavior regarding the object. In the absence of attitudes stored in memory, we would be required to appraise an object every time we encountered it to assess its evaluative implications and decide whether to approach the object or avoid it. This process would overwhelm our cognitive capacity and would severely limit our ability to act swiftly and decisively in situations that require immediate action.

The attitudes we hold sometimes serve other psychological functions as well. For example, some of our attitudes enable us to affirm central aspects of our self-concept by expressing our core values. Support for a particular affirmative action policy may enable an individual to express the central role that egalitarianism plays in his or her worldview. In this case, the policy attitude could be said to serve a "value-expressive" function. Other attitudes enable us to enjoy smooth social interactions with important others, serving a "social-adjustive" function. For example, holding a positive attitude toward environmental conservation may make it easier for us to get along with close friends who hold pro-environment attitudes. Still other attitudes serve an "ego-defensive" function, helping shield people from recognizing unpleasant aspects of themselves. For example, instead of acknowledging our own unacceptable impulses or feelings of inferiority, we may project these qualities onto out-groups. In this case, our negative attitudes toward the members enable us to distance ourselves from these negative qualities, protecting our self-image.

What Do Attitudes Do?

Attitudes are tremendously consequential. In fact, their influence can be detected almost immediately upon encountering an attitude object. Psychophysiological evidence reveals that almost instantly, the objects that we encounter are categorized according to our attitudes toward them—things that we like are differentiated from things that we dislike. This occurs even when we are not actively attending to the evaluative connotations of an object.

Once an attitude has been activated, it systematically influences thought and behavior. For example, attitudes often bias our judgments and shape our interpretations of events. This explains how supporters of two different political candidates can watch the very same debate and can come away convinced that his or her own candidate was clearly victorious. In this case, their pre-existing attitudes toward the candidates colored their interpretation of the debate performances.

And of course, attitudes motivate and guide behavior. For example, people's attitudes toward recycling are strongly predictive of whether or not they actually engage in recycling behavior. Attitudes toward particular consumer products powerfully shape people's purchasing decisions. And attitudes toward political candidates are excellent predictors of voting behavior.

Indeed, attitudes have been shown to predict behavior toward a diverse range of objects.

An Important Caveat

It is important to note, however, that attitudes do not always exert such powerful effects. In fact, attitudes sometimes appear to have negligible influence on thought and behavior. Recently, therefore, a central focus within the attitude literature has been on identifying the conditions under which attitudes do and do not powerfully regulate cognition and behavior. And indeed, great strides have been made in this effort.

It has been established, for example, that attitudes influence thought and behavior for some types of people more than others, and in some situations more than others. More recently, attitude researchers have determined that some attitudes are inherently more powerful than others. These attitudes profoundly influence our perceptions of and thoughts about the world around us, and they inspire us to act in attitude-congruent ways. Further, these attitudes tend to be tremendously durable, remaining stable across time and in the face of counter-attitudinal information. Other attitudes do not possess any of these qualities—they exert little influence on thought and behavior, they fluctuate over time, and they change in response to persuasive appeals.

The term *attitude strength* captures this distinction, and it provides important leverage for understanding and predicting the impact of attitudes on thought and behavior. That is, knowing an individual's attitude toward a particular object can be tremendously useful in predicting his or her behavior toward the object, but it is just as important to know the *strength* of the attitude.

Fortunately, several attitudinal properties have been identified that differentiate strong attitudes from weak ones, enabling scholars to measure these properties and draw inferences about the strength of a given attitude (and therefore about its likely impact on thought and behavior). For example, strong attitudes tend to be held with great certainty, based on a sizeable store of knowledge and on a good deal of prior thought, and considered personally important to the attitude holder. Thus, measures of attitude certainty, attitude-relevant knowledge, the extent of prior thought about the attitude object, and attitude importance offer valuable insights regarding the strength of individuals' attitudes.

Ambivalence is another important component of attitude strength. Sometimes people simultaneously experience both positive and negative reactions toward an object, producing an uncomfortable state of evaluative tension. Ambivalent attitudes tend to be weaker than univalent attitudes, so assessing ambivalence toward an attitude object can be very useful. Furthermore, on bipolar evaluative measures, people who have highly ambivalent attitudes often select the scale midpoint, rendering them indistinguishable from people who are neutral toward an object. Directly asking people how conflicted or how torn they feel about the attitude object or asking people for separate reports of their positivity and negativity toward the attitude object enable researchers to differentiate among these two groups of respondents.

Response latencies (i.e., the length of time it takes a person to answer an attitude question) can also reveal something about the strength of peoples' attitudes: attitudes that spring to mind and can be expressed quickly tend to be stronger than those that require deliberation. Increasingly, survey researchers have begun measuring the latency between the conclusion of an attitude question and the start of respondents' attitude response in an effort to capture differences in attitude accessibility. Because they do not involve additional survey items, response latencies have the potential to provide an efficient and cost-effective index of attitude strength. However, differences in survey response latency can be due to factors other than attitude accessibility. Furthermore, attitude accessibility is only one of several key strength-related attitude properties, and these properties are not always highly correlated. Thus, accessibility alone provides an imperfect index of attitude strength and whenever feasible, additional strength-related attitude properties (e.g., importance, certainty) should also be measured.

Asia A. Eaton and Penny S. Visser

See also Attitude Measurement; Attitude Strength; Bipolar Scale; Feeling Thermometer; National Election Studies (NES); Opinion Question; Opinions; Response Latency

Further Readings

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Fort Worth, TX: Harcourt Brace Jovanovich.

Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.),

The handbook of social psychology (4th ed., Vol. 1, pp. 269–322). New York: McGraw-Hill.
Fazio, R. H., & Olson, M. A. (2003). Attitudes: Foundations, functions, and consequences. In M. Hogg & J. Cooper (Eds.), *The SAGE handbook of social psychology.* London: Sage.
Petty, R. E., & Krosnick, J. A. (1995). *Attitude strength: Antecedents and consequences.* Mahwah, NJ: Lawrence Erlbaum.

ATTITUDE STRENGTH

Attitude strength refers to the extent to which an attitude is consequential. Compared to weak attitudes, strong attitudes are more likely to remain stable over time, resist influence, affect thought, and guide behavior.

Researchers have identified several attributes related to attitude strength. Several frequently studied attributes are well suited for survey research because they can be assessed directly using a single self-report survey item. For example, *attitude extremity* can be conceptualized as the absolute value of an attitude score reported on a bipolar scale that is centered at zero and ranges from strongly negative to strongly positive. *Attitude importance* is the significance people perceive a given attitude to have for them. *Attitude certainty* refers to how sure or how confident people are that their attitude is valid. Each of these attributes can be measured with straightforward questions, such as, *To what extent is your attitude about X positive or negative?; How important is X to you personally?;* and *How certain are you about your attitude about X?* Recent research suggests that attitude strength also is related to the extent that individuals subjectively associate an attitude with their personal moral convictions.

Other attributes can be assessed directly, with self-report survey items, or indirectly, with survey measures that allow researchers to infer the level of the attribute without relying on people's ability to introspect. For example, knowledge is the amount of information people associate with an attitude. Knowledge often is assessed by quizzes or by asking people to recall and list facts or experiences they relate to the attitude object. In a similar way, ambivalence, or the extent that people feel conflicted about a target, can be measured by asking people to list both positive and negative thoughts about the attitude object.

Most attitude strength research has assessed the association between attributes and characteristics of strong attitudes. Much less is known about how strength-related attributes relate to each other. Existing evidence, however, suggests that attitude attributes are best conceptualized as distinct constructs rather than as indicators of a single latent construct. Correlations between attributes typically range from low to only moderately positive.

Moreover, attributes often have different antecedents and consequences. For example, attitude importance, but not attitude certainty, about political policies has been found to predict whether people voted in the 1996 U.S. presidential election. In contrast, attitude certainty, but not attitude importance, has been found to predict whether people were willing to accept a nonpreferred candidate in the election.

Christopher W. Bauman

See also Attitude Measurement; Attitudes

Further Readings

Abelson, R. P. (1995). Attitude extremity. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 25–42). Mahwah, NJ: Lawrence Erlbaum.
Boninger, D. S., Krosnick, J. A., Berent, M. K., & Fabrigar, L. R. (1995). The causes and consequences of attitude importance. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 159–190). Mahwah, NJ: Lawrence Erlbaum.
Gross, R. A., Holtz, R., & Miller, N. (1995). Attitude certainty. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 215–246). Mahwah, NJ: Lawrence Erlbaum.
Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 1–24). Mahwah, NJ: Lawrence Erlbaum.
Raden, D. (1985). Strength-related attitude dimensions. *Social Psychology Quarterly*, *48*, 312–330.
Visser, P. S., Bizer, G. Y., & Krosnick, J. A. (2006). Exploring the latent structure of strength-related attitude attributes. *Advances in Experimental Social Psychology*, *39*, 1–67.

ATTRITION

Unit nonresponse is a problem for any type of survey; however, unit nonresponse in panel studies can be a more severe problem than in cross-sectional studies.

Like cross-sectional studies, panel studies are subject to nonresponse at the initial wave. In addition, attrition—which is unit nonresponse after the initial wave of data collection—can occur at each subsequent wave.

A framework for understanding attrition in panel studies divides the participation process into three conditional steps: (1) location, (2) contact given location, and (3) cooperation given contact; this process cycle is repeated at each wave. Attrition thus occurs because of a failure to relocate or recontact an eligible sample unit after the initial wave of data collection, and because of noncooperation (i.e., a refusal to participate again in the survey) or the inability to participate again.

The accumulation of attrition over several waves can substantially reduce the number of sample units, thereby reducing statistical power for any type of analysis, both cross-sectional and longitudinal. However, attrition may also introduce nonresponse bias in the survey estimates. Differential or selective attrition occurs when the characteristics of the sample units who drop out of the panel because of attrition differ systematically from the characteristics of sample units who are retained in the panel study.

Distinguishing between initial wave nonresponse and attrition is important because the reasons for attrition may be different from the reasons for nonresponse in the initial wave of a panel study or in cross-sectional studies, in general. Contrary to cross-sectional studies where sample units' judgments about participating in the survey are largely made during the brief interactions they have with survey interviewers when the request is formulated, sample units in panel studies with repeated survey requests and contacts in between data collection points have more information about the nature of the request being made and will be influenced by their personal survey experience in the initial wave or other previous waves. In addition, in the case of a panel study, and once the initial wave has been conducted, the interviewers are better informed than in the initial wave to select the best approach to successfully locate, contact, and convince sample units to participate in additional waves of the panel study.

There are two main strategies that survey researchers use to address attrition. The first is to reduce attrition rates by maximizing sample retention; the second is to develop post-survey adjustments to correct for the biasing effects of attrition. These two strategies are not mutually exclusive, and they often are used together.

The main goal of panel management or panel maintenance is to maintain participation of all sample members in the panel study after the initial wave. The specific techniques to reduce attrition in panel studies are focused on locating the sample unit and establishing sufficient rapport with the sample units to secure their continued participation. Panel studies can keep contact with the sample units and keep them interested in participating in the panel study by adopting a good panel maintenance plan and employing techniques of tracking and tracing. Acquiring detailed contact information, the organization of contact efforts, hiring skilled interviewers, and retaining staff over time are important components of a good panel maintenance plan. Tracking procedures aim to maintain contact with sample units in the period between waves in order to update addresses between interviews so that a current or more recent address is obtained for each sample unit prior to conducting the interview. Tracking procedures are adopted in an attempt to find the missing sample units and are used at the point of data collection when the interviewer makes his or her first call, discovers the sample member has moved, and tries to find a new address or telephone number.

The second approach to addressing attrition is to calculate adjustment weights to correct for possible attrition bias after the panel study has been conducted. Since nonresponse may occur at each successive wave of data collection, a sequence of nonresponse adjustments must be employed. A common procedure is first to compute adjustment weights for nonresponse in the initial wave. At Wave 2, the initial weights are adjusted to compensate for the sample units that dropped out because of attrition in Wave 2; at Wave 3, the Wave 2 weights are adjusted to compensate for the Wave 3 nonrespondents; and so on. Adjustment weighting is based on the use of auxiliary information available for both the sample units that are retained and the sample units that dropped out because of attrition. However, for the second and later waves of a panel study, the situation to find suitable auxiliary information is very different than in cross-sectional studies or in the initial wave because responses from the prior waves can be used in making the adjustments for nonresponse in subsequent waves.

Femke De Keulenaer

See also Differential Attrition; Nonresponse Bias; Nonresponse Rates; Panel; Panel Data Analysis; Panel Survey; Post-Survey Adjustments; Unit Nonresponse

Further Readings

Kalton, G., & Brick, M. (2000). Weighting in household panel surveys. In D. Rose (Ed.), *Researching social and economic change: The uses of household panel studies* (pp. 96–112). New York: Routledge.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (Eds.). (1989). *Panel surveys.* New York: Wiley.

Laurie, H., Smith, R., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, *15*, 269–282.

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves et al. (Eds.), *Survey nonresponse* (pp. 259–272). New York: Wiley-Interscience.

Audio Computer-Assisted Self-Interviewing (ACASI)

Audio computer-assisted self-interviewing (ACASI) is a methodology for collecting data that incorporates a recorded voice into a traditional computer-assisted self-interview (CASI). Respondents participating in an ACASI survey read questions on a computer screen and hear the text of the questions read to them through headphones. They then enter their answers directly into the computer either by using the keyboard or a touch screen, depending on the specific hardware used. While an interviewer is present during the interview, she or he does not know how the respondent answers the survey questions, or even which questions the respondent is being asked.

Typically the ACASI methodology is incorporated into a longer computer-assisted personal interview (CAPI). In these situations, an interviewer may begin the face-to-face interview by asking questions and recording the respondent's answers into the computer herself or himself. Then in preparation for the ACASI questions, the interviewer will show the respondent how to use the computer to enter his or her own answers. This training may consist solely of the interviewer providing verbal instructions and pointing to various features of the computer but could also include a set of practice questions that the respondent completes prior to beginning to answer the actual survey questions. Once the respondent is ready to begin answering the survey questions, the interviewer moves to a place where she or he can no longer see the computer screen but where she or he will still be able to answer questions or notice if the respondent appears to be having difficulties and to offer assistance as needed.

ACASI offers all the benefits of CASI, most notably: (a) the opportunity for a respondent to input her or his answers directly into a computer without having to speak them aloud to the interviewer (or risk having them overheard by someone else nearby); (b) the ability to present the questions in a standardized order across all respondents; (c) the ability to incorporate far more complex skip routing and question customization than is possible for a paper-based self-administered questionnaire; and (d) the opportunity to eliminate questions left blank, inconsistent responses, and out-of-range responses. In addition, the audio component allows semi-literate or fully illiterate respondents to participate in the interview with all of the same privacy protections afforded to literate respondents. This is significant, because historically, in self-administered surveys it was not uncommon for individuals who could not read to either be excluded from participation in the study altogether or to be included but interviewed in a traditional interviewer-administered manner, resulting in the potential for significant mode effects.

Evidence from several large-scale field experiments suggests the ACASI methodology reduces socially desirable responding compared to both interviewer-administered and solely text-based self-administration methods for sensitive topics, including use of illicit drugs, sexual behaviors, and abortion. ACASI also allows for increased standardization in the presentation of the survey questions because a pre-recorded voice is utilized to administer the survey questions. As a result, each respondent hears all introductory text, questions, and response categories read in exactly the same way. Thus, the natural variation caused by differences in interviewers' reading skills, pace, and/or vocal quality is eliminated.

Rachel Caspar

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Face-to-Face Interviewing; Interactive Voice Response (IVR); Mode Effects; Privacy; Self-Administered Questionnaire; Sensitive Topics; Social Desirability; Underreporting

Further Readings

O'Reilly, J., Hubbard, M. L., Lessler, J. T., Biemer, P. P., & Turner, C. F. (1994). Audio and video computer-assisted self-interviewing: Preliminary tests of new technologies

for data collection. *Journal of Official Statistics*, *10*, 197–214.
Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions. *Public Opinion Quarterly*, *60*, 275–321.
Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*, 867–873.

AURAL COMMUNICATION

Aural communication involves the transmission of information through the auditory sensory system—the system of speaking and hearing. It usually encompasses both verbal communication and paralinguistic communication to convey meaning. Aural communication can be used to transmit information independently or in combination with visual communication. When conducting surveys, the mode of data collection determines whether information can be transmitted aurally, visually, or both. Whether survey information is transmitted aurally or visually influences how respondents first perceive and then cognitively process information to provide their responses.

Aural communication relies heavily on verbal language when information is transmitted through spoken words. Additionally, paralinguistic or paraverbal communication, in which information is conveyed through the speaker's voice, is also an important part of aural communication. Paralinguistic communication can convey additional information through voice quality, tone, pitch, volume, inflection, pronunciation, and accent that can supplement or modify the meaning of verbal communication. Paralinguistic communication is an extremely important part of aural communication, especially in telephone surveys, where visual communication is absent.

Since aural and visual communication differ in how information is presented to survey respondents, the type of communication impacts how respondents initially perceive survey information. This initial step of perception influences how respondents cognitively process the survey in the remaining four steps (comprehension, retrieval, judgment formation, and reporting the answer). Whereas telephone surveys rely solely on aural communication, both face-to-face and Internet surveys can utilize aural and visual communication. Face-to-face surveys rely extensively on aural communication with the occasional use of visual communication by utilizing show cards or other visual aids. In contrast, Web surveys use mostly visual communication but have the potential to incorporate aural communication through sound files, a practice that is still fairly uncommon and generally only used to transmit information to respondents. Paper surveys do not utilize any aural communication.

The influence that aural communication has on perception and cognitive processing of information can contribute to effects between modes that rely primarily on aural communication and modes that rely primarily on visual communication. For example, aural transmission of information makes higher demands on memory capacity than visual transmission because respondents must remember information communicated to them without a visual stimulus to remind them. Additionally, in aural communication, the flow or pace is usually controlled by the interviewer, so the respondent may have more pressure to respond quickly rather than being able to fully process the information at his or her own pace. Because of these influences of aural communication on processing time and memory, surveyors often shorten questions and limit the amount of information respondents need to remember at one time in telephone surveys where aural communication cannot be supplemented by visual communication. However, this design difference can impact whether data from telephone surveys can be combined with or compared to data collected using primarily visual communication, where longer and more complex questions and sets of response options are often used.

Leah Melani Christian and Jolene D. Smyth

See also Mode Effects; Mode of Data Collection; Telephone Surveys; Visual Communication

Further Readings

de Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*, 233–255.
Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., II, & Waksberg, J. (Eds.). (1988). *Telephone survey methodology*. New York: Wiley.

AUXILIARY VARIABLE

In survey research, there are times when information is available on every unit in the population. If a variable that is known for every unit of the population is not a variable of interest but is instead employed to improve the sampling plan or to enhance estimation of the variables of interest, it is called an *auxiliary variable*.

Ratio and Regression Estimation

The term *auxiliary variables* is most commonly associated with the use of such variables, available for all units in the population, in ratio estimation, regression estimation, and extensions (calibration estimation).

The ratio estimator is a widely used estimator that takes advantage of an auxiliary variable to improve estimation. If x is the auxiliary variable and y is the variable of interest, let X and Y denote the population totals for x and y and let $\hat{X}$ and $\hat{Y}$ denote unbiased estimators of X and Y. Then the ratio estimator $\hat{Y}_R$ of Y is given by

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{X}}X.$$

$\hat{Y}_R$ improves upon $\hat{Y}$ provided that the correlation between x and y exceeds one-half of $S_x/\bar{X}$ divided by $S_y/\bar{Y}$ where $S_x, S_y, \bar{X}$, and $\bar{Y}$ are respectively the standard errors for x and y and the population means for x and y. The ratio estimator takes advantage of the correlation between x and y to well estimate Y/X by $\hat{Y}/\hat{X}$ and further takes advantage of X being known.

A more flexible estimator than the ratio estimator also taking advantage of the auxiliary variable x is the regression estimator:

$$\hat{Y}_{Reg} = \hat{Y} + \hat{b}(X - \hat{X}),$$

where $\hat{b}$ is the estimated slope of y on x from the sample data. The regression estimator can be extended to make use of a vector, X, of auxiliary variables rather than a single one.

In the case of stratified sampling, the ratio and regression estimators have a number of variants. In the case of ratio estimation, the separate ratio estimator does ratio estimation at the stratum level and then sums across strata, whereas the combined ratio estimator estimates $\hat{X}$ and $\hat{Y}$ across strata and then takes ratios.

Unequal Probability Sampling

In unequal probability sampling, the auxiliary variable x is termed a measure of size. The probability of selecting a unit is proportional to its measure of size. For example, in a survey of business establishments, the measure of size might be the number of employees or the total revenue of the establishment, depending on the purpose of the survey and the auxiliary information available. There are numerous sampling schemes for achieving selection probabilities proportional to the measure of size, one being unequal probability systematic sampling. Under general conditions, these schemes are more efficient than equal probability sampling when there is substantial variability in the size of the units in the population.

Stratification

It is often advantageous to divide a population into homogeneous groups called *strata* and to select a sample independently from each stratum. Auxiliary information on all population units is needed in order to form the strata. The auxiliary information can be a categorical variable (e.g., the county of the unit), in which case the categories or groups of categories form the strata. The auxiliary information could also be continuous, in which case *cut points* define the strata. For example, the income of a household or revenue of an establishment could be used to define strata by specifying the upper and lower limits of income or revenue for each stratum.

Post-Stratification

If specific auxiliary information is not used in forming strata or as a measure of size, it can still be used to adjust the sample weights to improve estimation in a process called *post-stratification.*

Michael P. Cohen

See also Bias; Imputation; Post-Stratification; Probability of Selection; Probability Proportional to Size (PPS) Sampling; Strata; Stratified Sampling; Systematic Sampling

Further Readings

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys.* New York: Wiley.
Levy, P. S., & Lemeshow, S. (1991). *Sampling of populations: Methods and applications.* New York: Wiley.
Särndal, C. E., Swennson, B., & Wretman, J. (1992). *Model-assisted survey sampling.* New York: Springer-Verlag.

B

BALANCED QUESTION

A balanced question is one that has a question stem that presents the respondent with both (all reasonably plausible) sides of an issue. The issue of "balance" in a survey question also can apply to the response alternatives that are presented to respondents. Balanced questions are generally closed-ended questions, but there is nothing inherently wrong with using open-ended questions in which the question stem is balanced.

For example, the following closed-ended question is *unbalanced* for several reasons and will lead to invalid (biased) data:

> Many people believe that American troops should be withdrawn from Iraq as soon as possible. Do you Strongly Agree, Agree, Somewhat Agree, or Strongly Disagree?

First, the question stem presents only one side of the issue in that it notes only one position taken by some people in the general public. Second, the response alternatives are not balanced (symmetrical), as there are three "agree" choices and only one extreme "disagree" choice. Third, the four response alternatives have no true midpoint; this is a further aspect of the asymmetrical (unbalanced) nature of the response alternatives.

In contrast, a balanced version of this question would be as follows:

> Some people believe that American troops should be withdrawn from Iraq as soon as possible, whereas other people believe that they should remain in Iraq until the country is more stable. What is your opinion on whether the troops should be withdrawn as soon as possible? Do you Strongly Agree, Somewhat Agree, Somewhat Disagree, or Strongly Disagree?

This wording is balanced because it poses both sides of the issue. It also has a symmetrical set of response alternatives, with two choices for "agree" and two similarly worded choices for "disagree." Furthermore, it has a true midpoint, even though that midpoint does not have an explicit response alternative associated with it. If the researchers wanted to add a fifth response option representing the midpoint, they could add, "Neither Agree nor Disagree" in the middle.

In writing survey questions, researchers can further balance them by using randomized variations of the ordering of the wording in the question stem and in the ordering of the response choices. In the second example presented here, one version of the stem could be worded as shown and a second version could have the information reversed, as in, *Some people believe that American troops should remain in Iraq until the country is more stable, whereas other people believe that they should be withdrawn from Iraq as soon as possible*. The response alternatives could also be randomly assigned to respondents so that some respondents received the four response choices shown in the second example, and the other half of the respondents could be presented with this order of response choices: Strongly Disagree, Somewhat Disagree, Somewhat Agree, or Strongly Agree.

Paul J. Lavrakas

See also Closed-Ended Question; Open-Ended Question; Question Stem; Random Assignment; Response Alternatives

Further Readings

AAPOR. (2007). *Question wording*. Retrieved March 11, 2008, from http://www.aapor.org/questionwording

Shaeffer, E. M., Krosnick, J. A., Langer, G. E., & Merkle, D. M. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly*, *69*(3), 417–428.

Balanced Repeated Replication (BRR)

Balanced repeated replication (BRR) is a technique for computing standard errors of survey estimates. It is a special form of the *replicate weights technique*. The basic form of BRR is for a stratified sample with two primary sampling units (PSUs) sampled with replacement in each stratum, although variations have been constructed for some other sample designs. BRR is attractive because it requires slightly less computational effect than the jackknife method for constructing replicate weights and it is valid for a wider range of statistics. In particular, BRR standard errors are valid for the median and other quantiles, whereas the jackknife method can give invalid results.

A sample with two PSUs in each stratum can be split into halves consisting of one PSU from each stratum. The PSU that is excluded from a half-sample is given weight zero, and the PSU that is included is given weight equal to 2 times its sampling weight. Under sampling with replacement or sampling from an infinite population, these two halves are independent stratified samples. Computing a statistic on each half and taking the square of the difference gives an unbiased estimate of the variance of the statistic. Averaging this estimate over many possible ways of choosing one PSU from each stratum gives a more precise estimate of the variance.

If the sample has L strata there are 2^L ways to take one PSU from each stratum, but this would be computationally prohibitive even for moderately large L. The same estimate of the variance of a population mean or population total can be obtained from a much smaller set of "splittings" as long as the following conditions are satisfied:

1. Each PSU is in the first half in exactly 50% of the splittings.
2. Any pair of PSUs from different strata is in the same half in exactly 50% of the splittings.

A set of replicates constructed in this way is said to be in *full orthogonal balance*. It is clearly necessary for these conditions that the number of splittings, R, is a multiple of 4.

An important open question in coding theory, the *Hadamard conjecture*, implies that a suitable set of splittings is possible whenever R is a multiple of 4 that is larger than L. Although the Hadamard conjecture is unproven, sets of replicates with full orthogonal balance are known for all values of R that are likely to be of interest in survey statistics. The construction is especially simple when R is a power of 2, which results in at most twice as many replicates as necessary.

All sets of replicates with full orthogonal balance give the same standard errors as the full set of 2^L replicates for the estimated population mean or population total, and thus it does not matter which set is chosen. For a statistic other than the mean or total, on the other hand, different sets of replicates in full orthogonal balance will typically not give exactly the same standard error. The difference is usually small, and analyses often do not report how the set of replicates was constructed.

One disadvantage of the BRR approach is that a half-sample increases the risk of small-sample computational difficulties such as zero cells in tables. A variant called Fay's method multiplies the sampling weights by $2 - \rho$ and ρ rather than 2 and 0, thus including all observations in all the computations. Fay's method retains the wide validity of BRR and has better small-sample performance. Fay's method is usually available in software that supports BRR replicate weights.

The other disadvantage of BRR is that it applies only to a specialized set of designs. This disadvantage is more difficult to avoid. There are variants of BRR that apply to designs for which the number of PSUs per stratum is fixed and small, but greater than 2. There are also variants that allow for a few strata to have extra or missing PSUs due to design imperfections. Methods for constructing these variants of BRR are typically not available in standard survey software.

Thomas Lumley

See also Jackknife Variance Estimation; Primary Sampling Unit (PSU); Replicate Methods for Variance Estimation; Standard Error; Stratified Sampling

Further Readings

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods* (pp. 212–217). Alexandria, VA: American Statistical Association.

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, *6*, 223–229.

Rao, J. N. K., & Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, *86*, 403–415.

Bandwagon and Underdog Effects

Bandwagon and underdog effects refer to the reactions that some voters have to the dissemination of information from trial heat questions in pre-election polls. Based upon the indication that one candidate is leading and the other trailing, a *bandwagon effect* indicates the tendency for some potential voters with low involvement in the election campaign to be attracted to the leader, while the *underdog effect* refers to the tendency for other potential voters to be attracted to the trailing candidate.

Background

Bandwagon and underdog effects were a concern of the earliest critics of public polls, and the founders of polling had to defend themselves against such effects from the start. The use of straw polls was common by the 1920s, and by 1935 a member of Congress had introduced an unsuccessful piece of legislation to limit them by constraining the use of the mails for surveys. A second piece of legislation was introduced in the U.S. Senate after the 1936 election, following on the heels of an editorial in *The New York Times* that raised concerns about bandwagon effects among the public as well as among legislators who saw poll results on new issues (even while the *Times* acknowledged such effects could not have been present in the 1936 election). A subsequent letter to the editor decried an "underdog" effect instead, and the debate was off and running.

In 1937, a scholarly article by Claude E. Robinson presented a defense of the polls that focused on two claims that he disputed empirically. One claim was that the release of the polling data depressed turnout; Robinson argued that turnout had steadily increased from 1924, when the straw polls came to prominence, until the 1936 election. And the second claim concerned the bandwagon effect. Robinson argued that it was too soon to judge that such an effect occurs, because the data did not show any clear demonstration of it; among the multiple instances he cited was the fact that in 1936 Republican candidate Alf Landon's support actually dropped after the release of the 1936 *Literary Digest* results showing Landon in the lead.

George Gallup and S. F. Rae, in 1940, addressed the issue just before the next presidential election, again citing empirical data from multiple states and discussing reactions to presidential candidates and issues in national surveys. They concluded that there were no demonstrable effects while holding out the possibility that additional research might produce evidence in the future. Their approach is interesting in that it discusses alternative research designs that could shed light on the phenomenon. One was the possibility of panel designs for surveys, and the other was the use of experiments, although they warned against using college students as subjects and of issues of external validity associated with unrealistic settings or issues to be evaluated.

The concepts themselves require some definition and specification in order to understand why research on their existence was limited and inconclusive for such a long time, allowing the public pollsters to defend themselves so well. Even when research designs became more refined, the magnitude of effects that could be demonstrated appeared to be relatively small, not enough to affect most elections but with the potential for an impact on close ones. In one sense, both bandwagon and underdog effects reflect a simple stimulus–response model. A potential voter has an initial predisposition, either toward a candidate or to abstain. After exposure to polling information disseminated through the media (newspapers and radio in the 1930s and all kinds of media now), the individual's preference shifts toward one or another candidate, based upon whether the candidate is leading or trailing in the polls. So the first implication of assessing such effects with a survey design is that there should be measurements of preferences over time, preferably with a panel design as suggested by Gallup and Rae. But such panel designs have rarely been present in survey research on underdog and bandwagon effects.

Limitations

A second consideration is that the likely size of the effects is small. This is due to the fact that as Election Day approaches and preferences crystallize, it is the strongest partisans who are most likely to participate. And their preferences are the most stable in the electorate. As a result, there is a relatively small proportion of the likely electorate, as opposed to the entire registered or voting age population, that could be subject to such effects. This implies that very large sample sizes are needed to detect such effects with confidence.

A third consideration is that these two effects do not occur in isolation, and as a result they may offset each other because they reflect responses in opposing directions. This represents another difficulty in searching for their occurrence in single cross-sectional surveys. This in fact was the main point of evidence and source of refutation of bandwagon and underdog effects used by the public pollsters in the early defense of their work. Given the historical record of accuracy of the major public pollsters, with an average deviation from the final election outcome of about 2 percentage points (excluding the 1948 election), the differences between final pre-election poll estimates at the national level and the popular vote for president have been very small.

It should also be noted that the full specification of models that predict candidate preference involve a large number of factors, a further complication for isolating published poll results as a cause. For all of these reasons, researchers interested in these phenomena turned to alternative designs involving variations on experiments. The experimental approach has a number of advantages, including isolating exposure to poll results as the central causal factor when randomization of subjects to various treatment groups and a control group is used to make all other things equal. An experimental design can also assess temporal order as well, verifying that candidate preference occurred (or changed) after exposure to the poll results. A well-designed experimental study will require many fewer subjects than the sample size for a survey-based design. At the same time, the kind of subjects used in many experiments, such as college undergraduates, can raise questions about the external validity of the results. And the nature of questioning and the kinds of stimuli used can as well.

Research

Michael Traugott's 1992 comprehensive review of research on bandwagon and underdog effects found mixed results, probably because the research designs suffered from many of the limitations previously discussed. Virtually all of the experiments were conducted with undergraduate students in a campus setting. They tend to demonstrate effects of exposure to information about the relative standing of candidates in polls, but the subjects were essentially new or beginning voters who tended not to have strong partisan attachments or a history of voting.

In one of the few surveys with a panel design, a 1976 study found that perceptions of the electorate's reactions to Gerald Ford and Jimmy Carter did have an effect on respondents' preferences, especially among those who were ambivalent about the candidates or uncertain of their own choices. Researchers who study the presidential nominating process focus on candidate "momentum" that builds during the primaries and caucuses, a particular form of a bandwagon effect that affects partisans rather than the general electorate. And a panel study conducted before and after Super Tuesday during this phase of the 1988 election showed that contagion was a more powerful explanation for growing support for George H. W. Bush than a desire to support the winner.

In a more elaborate panel conducted by Paul J. Lavrakas and his colleagues during the 1988 election campaign, which also included an imbedded experimental administration of question wordings, both underdog and bandwagon effects were observed. In a pre-election survey, a random half of the sample was given information about the current poll standing of George H. W. Bush and Michael Dukakis while a control group was not. There was an interaction of support levels for each candidate with level of education. Among those with less than a high school education, there was an increase in uncertainty about their preferences but no movement toward one candidate or the other. Among those with a high school education, there was no change in certainty about who they would vote for; but there was an underdog effect when exposed to the current poll standings showing Bush ahead of Dukakis. And those with the highest levels of education showed no change in certainty or candidate preference upon exposure to poll results.

A Canadian study with a similar design focused on two political issues rather than candidate choice, and it detected bandwagon effects of approximately 5 to 7 percentage points. This is the equivalent of conducting two experiments simultaneously, using abortion and Quebec sovereignty as the issues and a statement about poll results and the nature of change in them as stimuli; the bandwagon effect was present in each.

In conclusion, with additional attention devoted to specification of the bandwagon and underdog concepts and a deeper understanding of the conditions needed to demonstrate their presence, the results of recent research indicate that bandwagon and underdog effects can be produced under a variety of conditions. The strongest support for their presence comes from carefully designed experiments. While there may be issues of external validity associated with those conducted in the laboratory, those that are grounded in representative samples of adults or registered voters seem more compelling. The renewed interest in this area of the study of media effects, coupled with more sophisticated survey methodology, suggests that further research on this topic will be fruitful.

Michael Traugott

See also Election Polls; Experimental Design; External Validity; Media Polls; Panel Survey; Public Opinion; Straw Polls; Trial Heat Question

Further Readings

Gallup, G., & Rae, S. F. (1940). Is there a bandwagon vote? *Public Opinion Quarterly*, *4*, 244–249.

Kenney, P. J., & Rice, T. W. (1994). The psychology of political momentum. *Political Research Quarterly*, *47*, 923–938.

Lavrakas, P. J., Holley, J. K., & Miller, P. V. (1990). Public reactions to polling news during the 1988 presidential election campaign. In P. J. Lavrakas & J. K. Holley (Eds.), *Polling and presidential election coverage* (pp. 151–183). Newbury Park, CA: Sage.

Nadeau, R., Cloutier, E., & Guay, J.-H. (1993). New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review/Revue internationale de science politique*, *14*, 203–213.

Robinson, C. E. (1937). Recent developments in the straw-poll field-Part 2. *Public Opinion Quarterly*, *1*, 42–52.

Traugott, M. W. (1992). The impact of media polls on the public. In T. E. Mann & G. R. Orren (Eds.), *Media polls in American politics* (pp. 125–149). Washington, DC: Brookings Institution Press.

Behavioral Question

Behavioral questions are survey questions that ask about respondents' factual circumstances. They contrast with attitude questions, which ask about respondents' opinions. Typical behavioral questions target the respondent's household composition, sources of income, purchases, crime victimizations, hospitalizations, and many other autobiographical details. The Current Population Survey (CPS), for example, asks:

> *Have you worked at a job or business at any time during the past 12 months?*

Similarly, the National Crime Survey (NCS) includes the following behavioral item:

> *During the last 6 months, did anyone steal things that belonged to you from inside ANY car or truck, such as packages or clothing?*

Although these examples call for a simple "Yes" or "No" response, other behavioral items require dates (*When was the last time you . . .* ?), frequencies (*How many times during the last month did you . . .* ?), amounts (*How much did you pay for . . .* ?), and other data. The CPS and NCS examples concern the respondents' behavior in a loose sense, but other questions are less about behavior than about existing or past states of affairs. For example, the following question, from the National Health Interview Survey (NHIS), is more difficult to peg as a behavioral matter:

> *How much do you know about TB—a lot, some, a little, or nothing?*

For questions such as this, "factual question" may be a better label than "behavioral question."

Because behavioral questions often probe incidents in the respondents' pasts, such as jobs and burglaries, they place a premium on the respondents' memory of these incidents. Inability to recall relevant information is thus one factor that affects the accuracy of responses to such questions. Questions about events that took place long ago, that are unremarkable, or that can be confused with irrelevant ones are all subject to inaccuracy because of the burden they place on memory.

People's difficulty in recalling events, however, can lead them to adopt other strategies for answering behavioral questions. In deciding when an event happened, for example, respondents may estimate the time of occurrence using the date of a better-remembered neighboring event ("The burglary happened just after Thanksgiving; so it occurred about December 1"). In deciding how frequently a type of event happened, respondents may base their answer on generic information ("I usually go grocery shopping five times a month"), or they may remember a few incidents and extrapolate to the rest ("I went grocery shopping twice last week, so I probably went eight times last month"). These strategies can potentially compensate for recall problems, but they can also introduce error. In general, the accuracy of an answer to a behavioral question will depend jointly, and in potentially complex ways, on both recall and estimation.

Answers to behavioral questions, like those to attitude questions, can depend on details of question wording. Linguistic factors, including choice of words, grammatical complexity, and pragmatics, can affect respondents' understanding of the question and, in turn, the accuracy of their answers. Because behavioral questions sometimes probe frequencies or amounts, they can depend on the respondents' interpretation of adverbs of quantification, such as *usually, normally*, or *typically* (*How often do you usually/normally/typically go grocery shopping each month?*) or quantifiers of amounts, such as *a lot, some*, or *a little* (as in the NHIS example). Similarly, answers to these questions are a function of respondents' interpretation of the response alternatives. Respondents may assume, for example, that the response options reflect features of the population under study and base their response choice on this assumption.

Lance J. Rips

See also Measurement Error; Respondent-Related Error; Satisficing; Telescoping

Further Readings

Schwarz, N., & Sudman, S. (1994). *Autobiographical memory and the validity of retrospective reports*. New York: Springer-Verlag.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey responding*. Cambridge, UK: Cambridge University Press.

Behavioral Risk Factor Surveillance System (BRFSS)

The Behavioral Risk Factor Surveillance System (BRFSS) was developed in 1984 as a state-based system designed to measure behavioral risk factors associated with chronic diseases and some infectious diseases. The BRFSS is the world's largest ongoing, random-digit dialing telephone survey on health of adults ages 18 years or older. The survey is administered by the health departments in the 50 U.S. states, the District of Columbia, Puerto Rico, Guam, and the Virgin Islands. The target population is noninstitutionalized adults ages 18 years or older; however, BRFSS has also been used to collect information about children in the households. A large number of interviews (estimated at 350,000) are conducted annually, facilitating the development of local, state, and national estimates of health conditions and risk behaviors.

Participating areas use a standard core questionnaire of about 75 questions. In addition, states can elect to add their own questions or one or more optional standardized modules. In 2006, BRFSS offered 20 of these optional modules, which vary in number of questions and topic and averaged about six questions per module. The number of state-added questions also varies each year, with some states adding as many as 50. All information is self-reported. The core interview takes about 20 minutes to complete.

BRFSS data are collected by each state or territory with support from the Centers for Disease Control and Prevention (CDC). CDC helps to coordinate activities by the states and CDC-based programs, monitors and enforces standardized data collection protocols, ensures the validity and reliability of the data, assists the states in developing new methods and approaches to data collection, and provides BRFSS data files for public use. Because the states are responsible for conducting the survey, multiple contractors are involved. Standardization is achieved through the use of common training and interviewing protocols.

A stratified sample design is used, which facilitates production of estimates for 54 states and territories and for selected local areas. The Selected Metropolitan/Micropolitan Area Risk Trends (SMART-BRFSS) project uses BRFSS to develop estimates for selected metropolitan and micropolitan statistical areas (MMSAs) with 500 or more respondents. Data from the core survey in each state and territory are combined to produce national estimates.

BRFSS data are also used for rapid response surveillance during health emergencies. In the wake of the September 11, 2001, terrorist attacks in New York and Washington, D.C., BRFSS was used to monitor the mental health status of residents in the most affected areas. During the 2004–05 influenza season, when the supply of available influenza vaccine to the United States was cut nearly in half, the BRFSS was used to monitor influenza vaccination coverage during the season, providing national, state, and local health officials with critical information needed to make vaccine redistribution decisions and to inform public health messages encouraging vaccination among people in high-priority groups.

Procedures for maximizing response rates include online standardized interviewer training (required for all BRFSS interviewers), thorough pretesting of the survey questions, toll-free telephone numbers for participants, automated review of key quality indicators (e.g., response rates, refusal rates, percentage of key items with missing data, distribution of respondents by sex and age), and flexible calling schedules. BRFSS is conducted in English and Spanish.

New methodological approaches are tested extensively and regularly to ensure that the BRFSS continues to thrive as one of the leading public health surveillance systems in the world in the face of mounting technological, social, and legal barriers to telephone surveys. This research aims to (a) expand the utility of the surveillance system by developing special surveillance projects, including rapid response surveillance, follow-up surveys, and stand-alone surveillance; (b) identify, monitor, and address potential threats to the validity and reliability of BRFSS data (e.g., changes in telecommunications technologies, legal and privacy restrictions, and changes in social behaviors that might affect survey participation); and (c) develop and conduct innovative pilot studies designed to improve BRFSS's methods and to shape the future direction of the system (e.g., multiple modes of survey administration, address-based sampling, and on-phone interpreters to expand the number of languages in which BRFSS is offered). In addition, BRFSS is exploring the possibility of incorporating households that have only cell phones into the BRFSS sample and collecting physical measures from selected respondents to improve the accuracy of the survey estimates.

Strengths of the BRFSS include the high quality of state and local data, which are available for public health planning. The large state sample sizes, averaging 6,000 completed interviews per state annually, permit analysis of data on population subgroups within a state and development of local estimates for some areas. Data have been collected for many years, so trend data exist for each state or territory and for the nation. BRFSS also facilitates surveillance capacity building within a state or territory. BRFSS provides a basis on which states can develop and expand their data collection and analysis capabilities. The current BRFSS program extends beyond data collection to include a series of committees, workgroups, and conferences that are built around the surveillance effort to help to integrate national, state, and local programs.

Michael Link

Further Readings

Behavioral Risk Factor Surveillance System: http://www.cdc.gov/brfss

Link, M., Battaglia, M., Frankel, M., Osborn, L., & Mokdad, A. (2006). Address-based versus random-digit dialed surveys: Comparison of key health and risk indicators. *American Journal of Epidemiology*, *164*, 1019–1025.

Link, M., & Mokdad, A. (2005). Use of alternative modes for health surveillance surveys: Results from a web/mail/telephone experiment. *Epidemiology*, *16*, 701–704.

Mokdad, A., Stroup, D., & Giles, W. (2003). Public health surveillance for behavioral risk factors in a changing environment: Recommendations from the behavioral risk factor surveillance team. *MMWR Recommendations and Reports*, *52*(RR09), 1–12.

Behavior Coding

Behavior coding concerns the systematic assignment of codes to the overt behavior of interviewer and respondent in survey interviews. The method was developed by Charles Cannell and his colleagues at the University of Michigan in the 1970s. Behavior coding is a major tool used to evaluate interviewer performance and questionnaire design. Behavior coding is sometimes referred to as "interaction analysis," although interaction analysis is usually more specifically used in the sense of applying behavior coding to study the course of the interaction between interviewer and respondent.

The three main uses of behavior coding are (1) evaluating interviewer performance, (2) pretesting questionnaires, and (3) studying the course of the interaction between interviewer and respondent.

Evaluating Interviewer Performance

The use of behavior coding to evaluate interviewer performance primarily concerns how the interviewer reads scripted questions from the questionnaire. Typical codes include "Reads question correctly," "Reads question with minor change," "Reads question with major change," "Question incorrectly skipped," and "Suggestive probe." Usually the number of different codes for the purpose of evaluating interviewer performance ranges from five to 15.

Evaluating interviewer performance is usually part of the main field work. To this end, the interviews from the actual survey are audio-recorded. A sufficiently large sample of interviews from each interviewer is drawn (preferably 20 or more of each interviewer) and subjected to behavioral coding. Results may be in the form of "Interviewer X reads 17% of the questions with major change." These results are used to give the interviewer feedback, retrain him or her, or even withdraw him or her from the study.

Pretesting Questionnaires

If a particular question is often read incorrectly, this may be due to interviewer error, but it may also be a result of the wording of the question itself. Perhaps the question has a complex formulation or contains words that are easily misunderstood by the respondent. To prevent such misunderstandings, the interviewer may deliberately change the formulation of the question.

To gain more insight into the quality of the questions, the behavior of the respondent should be coded too. Typical codes for respondent behavior include "Asks repetition of the question," "Asks for clarification," "Provides uncodeable response" (e.g., "I watch television most of the days," instead of an exact number), or "Expresses doubt" (e.g., "About six I think, I'm not sure"). Most behavior coding studies use codes both for the respondent and the interviewer. The number of different codes may range between 10 and 20.

Unlike evaluating interviewer performance, pretesting questionnaires by means of behavioral coding requires a pilot study conducted prior to the main data collection. Such a pilot study should reflect the main study as closely as possible with respect to interviewers and respondents. At least 50 interviews are necessary, and even more if particular questions are asked less often because of skip patterns.

Compared to other methods of pretesting questionnaires, such as cognitive interviewing or focus groups, pretesting by means of behavior coding is relatively expensive. Moreover, it primarily points to problems rather than causes of problems. However, the results of behavior coding are more trustworthy, because the data are collected in a situation that mirrors the data collection of the main study. Moreover, problems that appear in the actual behavior of interviewer and respondent are real problems, whereas in other cases, for example in cognitive interviewing, respondents may report pseudo-problems with a question just to please the interviewer.

Interviewer–Respondent Interaction

If one codes both the behavior of interviewer and respondent and takes the order of the coded utterances into account, it becomes possible to study the course of the interaction. For example, one may observe from a pretesting study that a particular question yields a disproportionately high number of suggestive probes from the interviewer. Such an observation does not yield much insight into the causes of this high number. However, if one has ordered sequences of codes available, one may observe that these suggestive probes almost invariantly occur after an uncodeable response to that question. After studying the type of uncodeable response and the available response alternatives in more detail, the researcher may decide to adjust the formulation of the response alternatives in order to decrease the number of uncodeable responses, which in turn should decrease the number of suggestive probes.

In contrast, if the researcher merely looked at the sheer number of suggestive probings, he or she might have decided to adjust the interviewer training and warn the interviewers not to be suggestive, especially when asking the offending question. This may help a bit, but does not take away the cause of the problem.

As the previous example shows, interviewer–respondent interaction studies are focused on causes of particular behavior, that is, the preceding behavior of the other person. Because the researcher does not want to overlook particular causes, each and every utterance in the interaction is usually coded and described with some code. Hence, the number of different codes used in these studies can be quite high and exceeds 100 in some studies.

Behavior Coding Procedures

Recording Procedures

In a few cases, interviews are coded "live" (during the interview itself), sometimes by an observer, sometimes even by the interviewer herself. A main reason for live coding is that one does not need permission of the respondent to audio-record the interview. Another advantage is that results are quickly available, which can be especially useful in case of pretesting questionnaires.

In most studies, however, the interview is first audio-recorded. More recently, in the case of computer-assisted interviewing, the interview is recorded by the computer or laptop itself, thus eliminating the need for a separate tape recorder. Coding audio-recorded interviews is much more reliable than live coding, because the coder can listen repeatedly to ambiguous fragments.

If interviews are audio-recorded, they are sometimes first transcribed before coding. Transcripts yield more details than the codes alone. For example, if a particular question is often coded as "Read with major change," the availability of transcripts allows the researcher to look at the kind of mistakes made by the interviewer. Transcripts also make semi-automatic coding possible; a computer program can decide, for example, whether or not questions are read exactly as worded.

Full Versus Selective Coding

In interviewer-monitoring studies, it may be sufficient to code the utterances of the interviewer only; moreover, the researcher may confine himself to particular interviewer utterances, like question reading, probing, or providing clarification. Other types of utterances—for example, repeating the respondent's answer—are neglected. In pretesting studies, it is sometimes decided to code only behavior of the respondent. Also, in interaction studies, the researcher may use a form of such "selective" coding, neglecting all utterances after the answer of the respondent (e.g., if the respondent continues to elucidate the answer, this would not be coded). Alternatively, each and every utterance is coded. Especially in the case of interaction studies, this is the most common strategy.

All these procedural decisions have time and cost implications. Selective live coding is the fastest and cheapest, while full audio-recorded coding using transcriptions is the most tedious and costly but also yields the most information.

Wil Dijkstra

See also Cognitive Interviewing; Interviewer Monitoring; Questionnaire Design

Further Readings

Cannell, C. F., Lawson, S. A., & Hausser, D. L. (1975). *A technique for evaluating interviewer performance: A manual for coding and analyzing interviewer behavior from tape recordings of household interviews.* Ann Arbor: University of Michigan, Survey Research Center of the Institute for Social Research.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In. N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15–36). San Francisco: Jossey-Bass.

Ongena, Y. P., & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics*, *22*(3), 419–451.

BENEFICENCE

The National Research Act (Public Law 93348) of 1974 created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, which, among other duties, was charged with the responsibility of identifying, articulating, and fully explaining those basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects throughout the United States. The commission's findings have been detailed in a 1979 document typically referred to as "The Belmont Report" in recognition of the Smithsonian Institute satellite site where it was drafted, the Belmont Conference Center in Elkridge, Maryland. The Belmont Report identified three basic ethical principals for the conduct of research, and one of these is *beneficence*. (The other identified principles are *justice* and *respect for persons*.) The Belmont Report clearly states that the principle of beneficence has its roots in the long-standing ethical guidelines of the medical profession's Hippocratic Oath generally and, in particular, its maxims instructing physicians to "never do harm" while acting "according to [one's] ability and [one's] judgment."

From these ideas, three more fully articulated notions have been derived. First is the principle that researchers are obligated, not merely encouraged or expected, to take all reasonable steps to avoid inflicting foreseeable harm upon research participants. Second is that researchers are obligated to work toward maximizing the benefits that research subjects might experience from participation in a research program. This does not mean that it is required that a research program provide direct benefits to its research subjects, however. Similarly, investigators are obligated to attempt to maximize anticipated longer-term benefits that society or people in general might realize as a consequence of the study. Finally, beneficence incorporates the idea that exposing research participants to risk is justifiable. The reality that research is a human enterprise, one that relies upon the individual abilities and judgments of researchers acting within the frameworks of existing knowledge and cultural norms, is recognized. As such, it is ethically acceptable and permissible for research to possess or encompass potential for a protocol or well-meaning actions taken by an investigator to result in harm to participants; typically some level of risk is appropriate, and it is a judgment call as to what that risk level can and should be. To summarize, beneficence represents the process of balancing the trade-off between the potential benefits and the justifiable risk of potential harms associated with participation in research, and it is manifest in investigator efforts to minimize risks while maximizing potential benefits to the individual participant and/or society as a whole.

The term *risk* refers to both the likelihood of some type of harm being experienced by one or more research participants and the extent or severity of that harm in the event that harm is experienced. Therefore, assessments of the risks associated with a research project may take account of the combined probabilities and magnitudes of potential harms that might accrue to research participants. Furthermore, though one proclivity may be to think of harm as physical insults (such as pain, discomfort, injury, or toxic effects of drugs or other substances), the nature of potential harms can be wide and varied. Indeed, while the potential for physical harms typically is virtually nonexistent in survey research, other categories of potential harms frequently are relevant. These other categories include:

- Psychological and emotional harms (e.g., depression, anxiety, confusion, stress, guilt, embarrassment, or loss of self-esteem)
- Social or political harms (e.g., "labeling," stigmatization, loss of status, or discrimination in employment)
- Economic harms (e.g., incurring actual financial cost from participation), and
- Infringements of privacy or breaches of confidentiality (which, in turn, may result in psychological, emotional, social, political, or economic harms)

It is the principle of beneficence, along with the principles of justice and respect for human subjects, that stands as the foundation upon which the government-mandated rules for the conduct of research (Chapter 45, Subpart A, Section 46 of the Code of Federal Regulations) have been created under the auspices of the U.S. Department of Health and Human Services, Office of Human Research Protections.

Jonathan E. Brill

See also Confidentiality; Ethical Principles

Further Readings

U.S. Office of Human Research Protections: http://www.hhs.gov/ohrp/belmontArchive.html
U.S. Office of Human Subjects Research: http://ohsr.od.nih.gov/guidelines/belmont.html

Bias

Bias is a constant, systematic form or source of error, as opposed to *variance*, which is random, variable error. The nature and the extent of bias in survey measures is one of the most daunting problems that survey researchers face. How to quantify the presence of bias and how to reduce its occurrence are ever-present challenges in survey research. Bias can exist in myriad ways in survey statistics. In some cases its effect is so small as to render it ignorable. In other cases it is nonignorable and it can, and does, render survey statistics wholly invalid.

Overview

Survey researchers often rely upon estimates of population statistics of interest derived from sampling the relevant population and gathering data from that sample. To the extent the sample statistic differs from the true value of the population statistic, that difference is the error associated with the sample statistic. If the error of the sample statistic is systematic—that is, the errors from repeated samples using the same survey design do not balance each other out—the sample statistic is said to be biased. Bias is the difference between the average, or expected value, of the sample estimates and the target population's true value for the relevant statistic. If the sample statistic derived from an estimator is more often larger, in repeated samplings, than the target population's true value, then the sample statistic exhibits a *positive bias*. If the majority of the sample statistics from an estimator are smaller, in repeated samplings, than the target population's true value, then the sample statistic shows a *negative bias*.

Bias of a survey estimate differs from the error of a survey estimate because the bias of an estimate relates to the systematic and constant error the estimate exhibits in repeated samplings. In other words, simply drawing another sample using the same sample design does not attenuate the bias of the survey estimate. However, drawing another sample in the context of the error of a survey can impact the value of that error across samples.

Graphically, this can be represented by a bull's-eye in which the center of the bull's-eye is the true value of the relevant population statistic and the shots at the target represent the sample estimates of that population statistic. Each shot at the target represents an estimate of the true population value from a sample using the same survey design. For any given sample, the difference between the sample estimate (a shot at the target) and the true value of the population (the bull's-eye) is the error of the sample estimate.

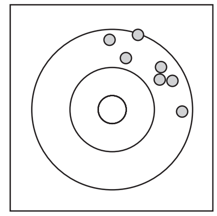

Figure 1 Example of a biased sample statistic

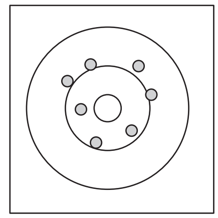

Figure 2 Example of an unbiased sample statistic

Multiple shots at the target are derived from repeated samplings using the same survey design. In each sample, if the estimator of the population statistic generates estimates (or hits on the bull's-eye) that are consistently off center of the target in a systematic way, then the sample statistic is biased.

Figure 1 illustrates estimates of the true value of the population statistic (the center of the bull's-eye), all of which are systematically to the upper right of the true value. The difference between any one of these estimates and the true value of the population statistic (the center of the bull's-eye) is the error of the estimate. The difference between the average value of these estimates and the center of the target (the true value of the population statistic) is the bias of the sample statistic.

Contrasting Figure 1 to a figure that illustrates an unbiased sample statistic, Figure 2 shows hits to the target that center around the true value, even though no sample estimate actually hits the true value.

Unlike Figure 1, however, the sample estimates in Figure 2 are not systematically off center. Put another way, the average, or expected value, of the sample estimates is equal to the true value of the population statistic indicating an unbiased estimator of the population statistic. This is an unbiased estimator even though all of the estimates from repeated samplings never hit the center of the bull's-eye. In other words, there is error associated with every sample estimate, but not bias.

Bias can be classified into two broad categories: (1) the bias related to the sampling process, and (2) the bias related to the data collection process. In the former case, if the survey design requires a sample to be taken from the target population, shortcomings in the sample design can lead to different forms of bias. Biases related to the sampling design are (a) estimation (or sampling) bias, (b) coverage bias, and (c) nonresponse bias. All of these are related to external validity.

Bias related to the data collection process is measurement bias and is related to construct validity. Measurement bias can be due to (a) data collection shortcomings dealing with the respondent, (b) the questionnaire, (c) the interviewer, (d) the mode of data collection, or (e) a combination of any of these.

To gauge the size of the bias, survey researchers sometimes refer to the *relative bias* of an estimator. The relative bias for an estimator is the bias as a proportion of the total population estimate.

Estimation Bias

Estimation bias, or sampling bias, is the difference between the expected value, or mean of the sampling distribution, of an estimator and the true value of the population statistic. More specifically, if θ is the population statistic of interest and $\hat{\theta}$ is the estimator of that statistic that is used to derive the sample estimate of the population statistic, the bias of $\hat{\theta}$ is defined as:

$$Bias[\hat{\theta}] = E[\hat{\theta}] - \theta.$$

The estimation bias of the estimator is the difference between the expected value of that statistic and the true value. If the expected value of the estimator, $\hat{\theta}$, is equal to the true value, then the estimator is unbiased.

Estimation bias is different from estimation, or sampling, error in that sampling error is the difference between a sample estimate and the true value of the population statistic based on one sampling of the sample frame. If a different sample were taken, using the same sample design, the sampling error would likely be different for a given sample statistic. However, the estimation bias of the sample statistic would still be the same, even in repeated samples.

Often, a desirable property of an estimator is that it is unbiased, but this must be weighed against other desirable properties that a survey researcher may want an estimator to have. For example, another desirable property of an estimator can be that it is the most efficient estimator from a class of estimators. In that case, even if the estimator is biased to some degree, the corresponding gain in efficiency can still lead to a smaller mean squared error when compared with unbiased estimators.

Coverage Bias

Coverage bias is the bias associated with the failure of the sampling frame to cover the target population. If the sampling frame does not allow the selection of some subset of the target population, then a survey can be susceptible to undercoverage. If a sampling frame enumerates multiple listings for a given member of the target population, then a survey can suffer from overcoverage.

In the case of undercoverage, a necessary condition for the existence of coverage bias is that there are members of the target population that are not part of the sampling frame. However, this is not a sufficient condition for coverage bias to exist. In addition, the members of the target population not covered by the sampling frame must differ across the population statistic of interest in some nonignorable way from the members of the target population covered by the sampling frame. To the extent that there is not a statistically significant nonignorable difference between the members of the target population covered by the sampling frame and the members of the target population not covered by the sampling frame, the coverage bias is likely to be small, even in instances when there is significant noncoverage of the population by the sampling frame.

If one defines the following:

$\bar{\theta}_C \equiv$ The population mean for the relevant variable for all members of the population covered by the sampling frame

$\bar{\theta}_{NC} \equiv$ The population mean for the relevant variable for all members of the population *not* covered by the sampling frame

$p_C \equiv$ The proportion of the target population covered by the sampling frame

coverage bias, due to undercoverage, is defined as:

$$Bias_{Coverage} \equiv (1 - p_C) * (\bar{\theta}_C - \bar{\theta}_{NC}).$$

Coverage bias is composed of two terms. The first term is the proportion of the target population not covered by the sampling frame. The second term is the difference in the relevant variable between the population mean for those members covered by the sampling frame and the population mean for those members not covered by the sampling frame. From this equation, it is clear that, as the coverage of the population by the sampling frame goes to 1, the amount of coverage bias goes to 0, even for large differences between the covered and noncovered population cohorts. Consequently, a sampling frame that covers the target population entirely cannot suffer from coverage bias due to undercoverage.

In those instances where there is not perfect overlap, however, between the target population and the sampling frame, methods have been developed to ameliorate possible coverage bias. Dual- and other multi-frame designs can be used to augment a single-frame design, thereby reducing the amount of noncoverage, which reduces the potential coverage bias. Another approach that can be used in conjunction with a dual-frame design is a mixed-mode survey, whereby different modes of data collection can be employed to address population members that would only be reached by one mode. Both of these approaches require implementation prior to data collection. However, post-survey weighting adjustments can be used, as the name implies, after data collection has taken place.

Nonresponse Bias

Nonresponse is the bias associated with the failure of members of the chosen sample to complete one or more questions from the questionnaire or the entire questionnaire itself. *Item nonresponse* involves sampled members of the target population who fail to respond to one or more survey questions. *Unit nonresponse* is the failure of sample members to respond to the entire survey. This can be due to respondents' refusals or inability to complete the survey or the failure of the researchers to contact the appropriate respondents to complete the survey.

Like coverage bias, to the extent that there is not a statistically significant nonignorable difference between the sample members who respond to the survey and the sample members who do not respond to the survey, the nonresponse bias is likely to be small (negligible), even in instances when there is significant item or unit nonresponse.

If one defines the following:

$\overline{\theta}_R \equiv$ The population mean for the relevant variable for all members of the sample who respond to the survey

$\overline{\theta}_{NR} \equiv$ The population mean for the relevant variable for all members of the sample who do *not* respond to the survey

$p_R \equiv$ The proportion of the sample that responds to the survey

nonresponse bias is defined as:

$$Bias_{Nonresponse} \equiv (1 - p_R) * (\overline{\theta}_R - \overline{\theta}_{NR}).$$

Nonresponse bias is composed of two terms. The first term is the proportion of the sample that did not respond to the survey (or to a question from the questionnaire in the case of item nonresponse). The second term is the difference in the relevant variable between the sample members who responded and the population mean for those sample members who did not respond. From this equation, it is clear that, as the response rate goes to 1, the amount of nonresponse bias goes to 0, even for large differences between the respondents and the nonrespondents. Consequently, a survey (or a question) that has a 100% response rate cannot suffer from nonresponse bias.

In those instances where there is not a 100% response rate, however, methods have been developed to lessen possible nonresponse bias. One method is to invest survey resources into maximizing the response rate to the survey. With this approach, regardless of how different respondents and nonrespondents might be, as the response rate goes to 1, the possibility of nonresponse bias may become more remote. However, often the survey resources required to achieve response rates that approach 100% are sizable. For example, in a telephone survey, conducting a large number of callbacks and undertaking refusal conversions can lead to higher response rates. But, by investing a large amount of the survey resources into higher response rates, the likelihood of diminished returns to this investment becomes more likely.

Survey researchers recognize that, in the context of nonresponse bias, the response rate is only part of the story. Therefore, some other methods that survey researchers use to combat nonresponse bias are (a) designing questionnaires that attempt to minimize the respondents' burden of completing the survey;

(b) identifying interviewers who are skilled in overcoming refusals and training these interviewers to hone these skills further; and (c) developing a motivational incentive system to coax reluctant respondents into participation.

Another approach that adjusts survey data to attempt to account for possible nonresponse bias is the use of post-stratified weighting methods, including the use of raking adjustments. With these methods, auxiliary information is used about the target population to bring the sample, along selected metrics, in line with that population. Imputation methods can also be used to insert specific responses to survey questions suffering from item nonresponse.

Measurement Bias

Measurement bias is the bias associated with the failure to measure accurately the intended variable or construct. The bias results from the difference between the true value for what the question or questionnaire intends to measure and what the question or questionnaire actually does measure. The source of the bias can be the interviewer, the questionnaire, the respondent, the mode of data collection, or a combination of all of these.

Measurement bias can be particularly difficult to detect. The problem with detection stems from the possibility that the bias can originate from so many possible sources. Respondents can contribute to measurement bias due to limitations in cognitive ability, including recall ability, and due to motivational shortcomings in the effort required to answer the survey questions properly. To combat measurement bias from respondents, surveys can be designed with subtle redundancy in the questions asked for variables and constructs where the survey researcher suspects some problem. This redundancy allows the researcher to examine the survey results for each respondent to determine whether internal inconsistencies exist that would undermine the data integrity for a given respondent.

The questionnaire can contribute to measurement bias by having questions that inadequately address or measure the concepts, constructs, and opinions that make up the subject matter of the study. The questionnaire can also contribute to measurement bias if the question wording and order of questions impact the quality of respondents' answers. Typically, the amount of measurement bias introduced due to the questionnaire will be difficult to gauge without controlled experiments to measure the difference in respondents' answers from the original questionnaire when compared to the questionnaire that was reworded and that reordered questions and possible response options.

Interviewers can contribute to measurement error by failing to read survey questions correctly, by using intonations and mannerisms that can influence respondents' answers, and by incorrectly recording responses. To address possible measurement bias from interviewers, the researcher can invest additional survey resources into the training of interviewers to eliminate habits and flawed data collection approaches that could introduce measurement bias. Moreover, the researcher can focus efforts to monitor interviewers as data collection is taking place to determine whether measurement bias is likely being introduced into the survey by interviewers.

The mode of data collection can also contribute to measurement bias. To the extent that respondents' answers are different across different modes of data collection, even when other factors are held constant, measurement bias could result due to different data collection modes.

Jeffery A. Stec

See also Construct Validity; Coverage Error; Dual-Frame Sampling; External Validity; Ignorable Nonresponse; Imputation; Interviewer Monitoring; Interviewer Training; Mean Square Error; Measurement Error; Missing Data; Mixed-Mode; Mode of Data Collection; Multi-Frame Sampling; Nonignorable Nonresponse; Nonresponse Error; Overcoverage; Post-Stratification; Questionnaire Design; Raking; Random Error; Sample Design; Sampling Error; Systematic Error; Target Population; True Value; Undercoverage; Unit Nonresponse; Variance

Further Readings

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. Toronto, Ontario, Canada: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Bilingual Interviewing

Bilingual interviewing refers to in-person and telephone surveys that employ interviewers who have the ability to speak more than one language. Typically in the United States, this means they are fluent in English and in Spanish. These interviewers use their language abilities to gain cooperation from sampled respondents and/or to gather data from these respondents.

It has become increasingly common for survey research organizations and their clients to gather the voices, viewpoints, and experiences of respondents who speak only in a native language other than English or prefer to speak in a language other than English. Representation from a sample that closely resembles the target population is important in reducing possible coverage and nonresponse biases. Even though the most common bilingual ethnic group in the United States is the Spanish-speaking or "Spanish Dominant" group, some survey researchers have been known to delve deep into ethnic communities, collecting survey data in more than 10 languages.

Knowing the Population

Bilingual interviewing presents a number of considerations for the survey researcher. First, survey researchers and clients need to determine which bilingual and non-English populations will be included in the survey. Before the questionnaire is translated into the foreign language(s), it is important to understand the bilingual population the survey will reach. Some bilingual populations have cultural perceptions about survey research that are different from non-bilingual populations. Foreign-born bilingual respondents often are not familiar with the field and practice of survey research, necessitating an easily understood explanation of the purpose of the survey provided by the interviewer at the time of recruitment, thereby increasing the level of trust between the interviewer and respondent.

Interviewer Support

Additionally, bilingual populations may show hesitation in answering particular questions that may not be problematic for non-bilingual populations. For example, many Spanish-speaking respondents tend to routinely hesitate when asked to provide their names and addresses. Each bilingual group may have its own set of questions that are considered "sensitive" when asked by an outsider (i.e., the survey interviewer). Thus the interviewer will need to find ways to minimize respondent hesitation and reluctance in order to continue successfully with the questionnaire. In order to anticipate sensitive questions, the researcher may want to hold focus groups with members of the bilingual population prior to the start of the study. Alterations to wording, improvements to transitions leading into question sequences, clarifying statements, and the addition of proactive persuaders can be useful in minimizing the negative effects of asking sensitive survey questions in languages other than English. The training bilingual interviewers receive thus needs to include attention to all these matters.

The survey researcher also will want to find out how the target population might respond to the survey mode. Some bilingual populations prefer to be interviewed in person, where they can see the facial expressions of the interviewer and pick up on body language. Other bilingual populations are more private and may prefer to be interviewed over the phone. Even though each bilingual population might have its own preference, the client and researchers may choose to use only one mode of data collection across different or mixed ethnic groups. Survey researchers can train bilingual interviewers on techniques to make the bilingual respondent feel comfortable in any type of survey mode.

Translation Process

The quality of the bilingual questionnaire translation will depend on the time and resources the survey researcher can devote to the task. It is in the best interest of the survey researcher to provide the group that is doing the translation with information on the background of the study, information about the questionnaire topics, country-of-origin statistics of the target population, acculturation level of the target population, effective words or phrases that may have been used in prior studies, and the format in which the survey will be conducted (i.e., phone, mail, in person, etc.). All of this information provides the translators with the tools to tailor the questionnaire translation to the bilingual target population(s). The preferred method of translation is to allow at least two translators to independently develop their own translated versions of the survey questionnaire. Next, the two translators use their independent versions to develop a single version and review the new version with the project lead to make sure the concepts have been conveyed correctly and effectively. The team then finalizes the version for use in bilingual interviewing pilot testing. Even though this translation process takes additional time and resources, it is preferred as a way to avoid problems common in most survey translations that are associated with (a) the overreliance of word-for-word literal translations, (b) oral surveys that are translated into written style (vs. spoken style), (c) translations in which the educational level is too high for the average respondent, (d) terms that do not effectively convey the correct meaning in the non-English language, (e) terms that are misunderstood, and (f) terms that are inappropriate to use in a professional survey. These problems become evident when the survey researcher has not provided enough information to the translation group.

The survey researcher will want to conduct the final check of translated document for words that may not be appropriate to use with the targeted bilingual population(s). Word meaning can vary by country, culture, and regional dialect, and inappropriate meanings may not be evident to the translation company. It is helpful to have a staff member who is knowledgeable in both bilingual translations and cultural considerations conduct the final questionnaire review. A fine-tuned script is essential to building trust and rapport with the bilingual respondent and to avoid any fear or hesitation invoked by an outside party collecting personal information.

Interviewing

In order to interview bilingual populations, the survey research organization must employ bilingual interviewers and bilingual support staff that are fluent in all the languages in which respondents will be recruited and interviewed. Interviewers and support staff should be able to show mastery of the relevant languages, and their abilities (including their ability to speak English or the dominant language in which the survey will be administered) should be evaluated through use of a language skills test to measure spoken fluency, reading ability, and comprehension in the other language(s). During data collection, it is important for interviewers and support staff to be able to communicate with the researchers and project supervisors to work together to address any culturally specific problem that may arise.

Depending on the level of funding available to the survey organization, there are a few areas of additional training that are useful in improving bilingual staff interviewing skills: listening techniques, language and cultural information about bilingual respondents, and accent reduction techniques.

The researcher may want to have bilingual interviewers trained to listen for important cues from the respondent, that is, the respondents' dominant language, level of acculturation, culture or country of origin, immigration status, gender, age, education level, socioeconomic status, individual personality, and situation or mood. The bilingual interviewer can use these cues proactively to tailor the survey introduction and address any respondent concerns, leading to a smooth and complete interview.

Survey researchers can provide interviewers with information on language patterns, cultural concepts, and cultural tendencies of bilingual respondents. Understanding communication behavior and attitudes can also be helpful in tailoring the introduction and addressing respondent concerns. Survey researchers need to train bilingual interviewers to use a "standard" conversational form of the foreign language, remain neutral, and communicate in a professional public-speaking voice. The use of a professional voice helps reduce the tendency of both the interviewer and respondent to judge social characteristics of speech, especially when the interviewer has the same regional language style as the respondent.

For those bilingual interviewers who will also be conducting interviews in English but have trouble with English consonant and vowel pronunciation, a training module that teaches *accent reduction* will help the interviewer produce clearer speech so that English-language respondents do not have to strain to understand.

Kimberly Brown

See also Fallback Statements; Interviewer Debriefing; Interviewer Training; Language Barrier; Language Translations; Nonresponse Bias; Questionnaire Design; Respondent–Interviewer Rapport; Sensitive Topics

Further Readings

Harkness, J. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: Wiley.

Harkness, J., Pennell, B. E., & Schoua-Glusberg, A. (2004). Questionnaire translation and assessment. in S. Presser, J. Rothgeb, M. Couper, J. Lessler, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Hoboken, NJ: Wiley.

Harkness, J., & Schoua-Glusberg, A. (1998). Questionnaires in translation. In J. Harkness (Ed.), *Cross-cultural survey* *equivalence* (pp. 87–127). ZUMA-Nachrichten Special no. 3. Mannheim: ZUMA.

Schoua-Glusberg, A. (1992). *Report on the translation of the questionnaire for the National Treatment Improvement Evaluation Study*. Chicago: National Opinion Research Center.

Schoua-Glusberg, A. (1998, May). *A focus-group approach to translating questionnaire items*. Paper presented at the 43rd Annual Meeting of the American Association for Public Opinion Research, Toronto, Ontario, Canada.

U.S. Census Bureau. (2004, April). *Census Bureau guideline: Language translation of data collection instruments and supporting materials*. Retrieved March 12, 2008, from http://www.census.gov/cac/www/007585.html

Bipolar Scale

Survey researchers frequently employ rating scales to assess attitudes, behaviors, and other phenomena having a dimensional quality. A rating scale is a response format in which the respondent registers his or her position along a continuum of values. The *bipolar scale* is a particular type of rating scale characterized by a continuum between two opposite end points. A central property of the bipolar scale is that it measures both the direction (side of the scale) and intensity (distance from the center) of the respondent's position on the concept of interest.

The construction of bipolar scales involves numerous design decisions, each of which may influence how respondents interpret the question and identify their placement along the continuum. Scales typically feature equally spaced gradients between labeled end points. Data quality tends to be higher when all of the gradients are assigned verbal labels than when some or all gradients have only numeric labels or are unlabeled. Studies that scale adverbial expressions of intensity, amount, and likelihood may inform the researcher's choice of verbal labels that define relatively equidistant categories.

Both numeric and verbal labels convey information to the respondent about the meaning of the scale points. As shown in Figure 1, negative-to-positive numbering (e.g., –3 to + 3) may indicate a bipolar conceptualization with the middle value (0) as a balance point. By contrast, low-to-high positive numbering (e.g., 0 to + 7) may indicate a unipolar conceptualization, whereby the low end represents the absence of the concept of interest and the high end represents a great deal. The choice of gradient labels may either reinforce or dilute the implications of the end point labels.

Extremely dissatisfied	–3	–2	–1	0	+1	+2	+3	Extremely satisfied

Figure 1 Example of bipolar scale

While negative-to-positive numbering may seem the natural choice for a bipolar scale, this format has a potential drawback. In general, respondents are less likely to select negative values on a scale with negative-to-positive labeling than they are to select the formally equivalent values on a scale with low-to-high positive labeling. Similarly, bipolar verbal labels result in more use of the midpoint and less use of the negative values than when unipolar verbal labels are used. Systematic reluctance to select negative values shifts the distribution of the responses to the positive end of the scale, yielding a relatively high mean score. In addition, the spread of the responses attenuates, yielding a reduction in variance.

The number of gradients represents a compromise between the researcher's desire to obtain more detailed information and the limited capacity of respondents to reliably make distinctions between numerous scale values. Research suggests that 7-point scales tend to be optimal in terms of reliability (test–retest) and the percentage of undecided respondents. Thus, 7-point scales plus or minus 2 points are the most widely used in practice.

Scales featuring a large number of labeled gradients may be difficult to administer aurally, as in a telephone interview. A common solution is to decompose the scale into two parts through a process called "branching" or "unfolding." The respondent is first asked about direction (e.g., *Overall, are you satisfied or dissatisfied?*) and then about degree (e.g., *Are you extremely (dis)satisfied, very (dis)satisfied, somewhat (dis)satisfied, or only a little (dis)satisfied?*). In certain multi-mode studies, branching may also be used to increase the comparability of responses across different modes of administration. In self-administered modes and face-to-face interviewing, respondents are often provided with a pictorial rendering of the scale, but respondents in telephone interviews usually cannot be provided with such visual aids. Administering a common branching question in each mode reduces the effect of mode on respondents' answers.

The midpoint of a bipolar scale may be interpreted in different ways. It can be conceived of as signaling indifference (e.g., neither satisfied nor dissatisfied) or ambivalence (e.g., satisfied in some ways but dissatisfied in others). When a middle position is explicitly offered, more respondents will select it than will volunteer it if it is not explicitly offered. In general, including a midpoint reduces the amount of random measurement error without affecting validity. If, however, the researcher has a substantive interest in dichotomizing respondents between the two poles, excluding a middle position may simplify the analysis.

Courtney Kennedy

See also Attitude Measurement; Branching; Guttman Scale; Likert Scale; Rating; Semantic Differential Technique; Questionnaire Design; Unfolding Question

Further Readings

Alwin, D. F., & Krosnick, J.A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods and Research*, *20*, 139–181.

Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology: Vol. I. Theory and method* (pp. 509–566). New York: Random House.

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570–582.

Bogus Question

A bogus question (also called a *fictitious question*) is one that asks about something that does not exist. It is included in a survey questionnaire to help the researcher estimate the extent to which respondents are providing ostensibly substantive answers to questions they cannot know anything about, because it does not exist. Bogus questions are a valuable way for researchers to gather information to help understand the nature and size of respondent-related measurement error.

Examples of how a researcher can use a bogus question abound, but they are especially relevant to surveys that measure recognition of, or past experience with, people, places, or things. For example, in pre-election polls at the time of the primaries, candidate name recognition is critical for understanding the intentions of voters. Thus, the name of a fictitious candidate could be added to the list of real candidates the survey is asking about to learn how many respondents answer that they know the fictitious (bogus) candidate. Similarly, when people (e.g., surveys of teenagers) are asked about the use of illegal substances they may have used in the past, it is advisable to add one or more bogus substances to the list of those asked about to be able to estimate the proportion of respondents who may well be answering erroneously to the real survey questions.

Past experience has shown that in some cases as many as 20% of respondents answer affirmatively when asked if they ever have "heard about X before today," where X is something that does not exist. That is, these respondents do not merely answer that they are "uncertain"—they actually report, "Yes," they have heard of the entity being asked about. Past research has suggested that respondents with lower educational attainment are most likely to answer affirmatively to bogus questions.

The data from bogus questions, especially if several bogus questions are included in the questionnaire, can be used by researchers to (a) filter out respondents who appear to have answered wholly unreliably, and/or (b) create a scaled variable based on the answers given to the bogus questions and then use this variable as a covariate in other analyses. Researchers need to explicitly determine whether or not the needs of the survey justify the costs of adding bogus questions to a questionnaire. When a new topic is being studied—that is, one that people are not likely to know much about—it is especially prudent to consider the use of bogus questions.

Paul J. Lavrakas

See also Measurement Error; Respondent-Related Error

Further Readings

Allen, I. L. (1966). Detecting respondents who fake and confuse information about question areas on surveys. *Journal of Applied Psychology*, *50*(6), 523–528.

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*, *50*(2), 240–250.

Lavrakas, P. J., & Merkle, D. M. (1990, November). *Name recognition and pre-primary poll measurement error*. Paper presented at International Conference of Survey Measurement Error, Tucson, AZ.

BOOTSTRAPPING

Bootstrapping is a computer-intensive, nonparametric approach to statistical inference. Rather than making assumptions about the sampling distribution of a statistic, bootstrapping uses the variability within a sample to estimate that sampling distribution empirically. This is done by randomly resampling with replacement from the sample many times in a way that mimics the original sampling scheme. There are various approaches to constructing confidence intervals with this estimated sampling distribution that can be then used to make statistical inferences.

Goal

The goal of statistical inference is to make probability statements about a population parameter, θ, from a statistic, $\hat{\theta}$, calculated from sample data drawn randomly from a population. At the heart of such analysis is the statistic's sampling distribution, which is the range of values it could take on in a random sample of a given size from a given population and the probabilities associated with those values. In the standard parametric inferential statistics that social scientists learn in graduate school (with the ubiquitous *t*-tests and *p*-values), a statistic's sampling distribution is derived using basic assumptions and mathematical analysis. For example, the central limit theorem gives one good reason to believe that the sampling distribution of a sample mean is normal in shape, with an expected value of the population mean and a standard deviation of approximately the standard deviation of the variable in the population divided by the square root of the sample size. However, there are situations in which either no such parametric statistical theory exists for a statistic or the assumptions needed to apply it do not hold. In analyzing survey data, even using well-known statistics, the latter problem may arise. In these cases, one may be able to use bootstrapping to make a probability-based inference to the population parameter.

Procedure

Bootstrapping is a general approach to statistical inference that can be applied to virtually any statistic. The basic procedure has two steps: (1) estimating the statistic's sampling distribution through resampling, and (2) using this estimated sampling distribution to construct confidence intervals to make inferences to population parameters.

Resampling

First, a statistic's sampling distribution is estimated by treating the sample as the population and conducting a form of Monte Carlo simulation on it. This is done by randomly resampling with *replacement* a large number of samples of size *n* from the original sample of size *n*. Replacement sampling causes the resamples to be similar to, but slightly different from, the original sample, because an individual case in the original sample may appear once, more than once, or not at all in any given resample.

For the resulting estimate of the statistic's sampling distribution to be unbiased, resampling needs to be conducted to mimic the sampling process that generated the original sample. Any stratification, weighting, clustering, stages, and so forth used to draw the original sample need to be used to draw each resample. In this way, the random variation that was introduced into the original sample will be introduced into the resamples in a similar fashion. The ability to make inferences from complex random samples is one of the important advantages of bootstrapping over parametric inference. In addition to mimicking the original sampling procedure, resampling ought to be conducted only on the random component of a statistical model. For example, an analyst would resample the error term of a regression model to make inferences about regression parameters, as needed, unless the data are all drawn from the same source, as in the case of using data from a single survey as both the dependent and independent variables in a model. In such a case, since the independent variables have the same source of randomness—an error as the dependent variable—the proper approach is to resample whole cases of data.

For each resample, one calculates the sample statistic to be used in the inference, $\hat{\theta}^*$. Because each resample is slightly and randomly different from each other resample, these $\hat{\theta}^*$s will also be slightly and randomly different from one another. The central assertion of bootstrapping is that a relative frequency distribution of these $\hat{\theta}^*$s is an unbiased estimate of the sampling distribution of $\hat{\theta}$, given the sampling procedure used to derive the original sample being mimicked in the resampling procedure.

To illustrate the effect of resampling, consider the simple example in Table 1. The original sample was drawn as a simple random sample from a standard normal distribution. The estimated mean and standard deviation vary somewhat from the population parameters (0 and 1, respectively) because this is a random sample. Note several things about the three resamples. First, there are no values in these resamples that do not appear in the original sample, because these resamples were generated from the original sample. Second, due to resampling with replacement, not every value in the original sample is found in each resample, and some of the original sample values are found in a given resample more than once. Third, the sample statistics estimated from the resamples (in this case, the means and standard deviations) are close to, but slightly different from, those of the original sample. The relative frequency distribution of these means (or standard deviations or any other statistic calculated from these resamples) is the bootstrap estimate of the sampling distribution of the population parameter.

How many of these resamples and $\hat{\theta}^*$s are needed for an analyst to conduct valid bootstrap inference? This bootstrap estimate of the sampling distribution of $\hat{\theta}$ is asymptotically unbiased, but how many resamples yield a sampling distribution estimate with a variance small enough to yield inferences precise enough to be practical? There are two components to this answer. First, the asymptotics of the unbiasedness proof for the bootstrap estimate of the sampling distribution require an original sample of data so that the statistical estimate has about 30–50 degrees of freedom. That is, bootstrapping needs samples of only about 30–50 cases more than the number of parameters being estimated. Second, the number of resamples needed to flesh out the estimated sampling distribution needs to be at least about 1,000. But with high-powered personal computers, such resampling and calculation requires a trivial amount of time and effort, given the ability to write an appropriate looping algorithm.

Confidence Intervals

After one estimates the sampling distribution of $\hat{\theta}$ with this resampling technique, the next step in bootstrap statistical inference is to use this estimate to construct confidence intervals. There are several ways to do this, and there has been some controversy as

Table 1 Original data and three resamples

Case Number	*Original Sample (N(0,1))*	*Resample #1*	*Resample #2*	*Resample #3*
1	0.697	−0.27	−1.768	−0.27
2	−1.395	0.697	−0.152	−0.152
3	1.408	−1.768	−0.27	−1.779
4	0.875	0.697	−0.133	2.204
5	−2.039	−0.133	−1.395	0.875
6	−0.727	0.587	0.587	−0.914
7	−0.366	−0.016	−1.234	−1.779
8	2.204	0.179	−0.152	−2.039
9	0.179	0.714	−1.395	2.204
10	0.261	0.714	1.099	−0.366
11	1.099	−0.097	−1.121	0.875
12	−0.787	−2.039	−0.787	−0.457
13	−0.097	−1.768	−0.016	−1.121
14	−1.779	−0.101	0.739	−0.016
15	−0.152	1.099	−1.395	−0.27
16	−1.768	−0.727	−1.415	−0.914
17	−0.016	−1.121	−0.097	−0.860
18	0.587	−0.097	−0.101	−0.914
19	−0.27	2.204	−1.779	−0.457
20	−0.101	0.875	−1.121	0.697
21	−1.415	−0.016	−0.101	0.179
22	−0.860	−0.727	−0.914	−0.366
23	−1.234	1.408	−2.039	0.875
24	−0.457	2.204	−0.366	−1.395
25	−0.133	−1.779	2.204	−1.234
26	−1.583	−1.415	−0.016	−1.121
27	−0.914	−0.860	−0.457	1.408
28	−1.121	−0.860	2.204	0.261
29	0.739	−1.121	−0.133	−1.583
30	0.714	−0.101	0.697	−2.039

Table 1 (continued)

Case Number	Original Sample (N(0,1))	Resample #1	Resample #2	Resample #3
Mean	−0.282	−0.121	−0.361	−0.349
StDev	1.039	1.120	1.062	1.147

Note: Column 2 holds the original sample of 30 cases drawn randomly from a standard normal distribution. Columns 3–5 hold bootstrap re-samples from the original sample.

to which confidence interval approach is the most practical and statistically correct. Indeed, much of the discussion of the bootstrap in the statistical literature since its development in the 1980s has been devoted to developing and testing these confidence interval approaches, which are too complicated to discuss here. (*See* Further Readings for details and instructions on these confidence interval approaches.)

Useful Situations

There are two situations in which bootstrapping is most likely to be useful to social scientists. First, the bootstrap may be useful when making inferences using a statistic that has no strong parametric theory associated with it, such as the indirect effects of path models, eigenvalues, the switch point in a switching regression, or the difference between two medians. Second, the bootstrap may be useful for a statistic that may have strong parametric theory under certain conditions, but those conditions do not hold. Thus, the bootstrap may be useful as a check on the robustness of parametric statistical tests in the face of assumption violations.

Christopher Z. Mooney

See also Confidence Interval; Dependent Variable; Independent Variable; Relative Frequency; Simple Random Sample

Further Readings

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley-Interscience.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

BOUNDING

Bounding is a technique used in panel surveys to reduce the effect of telescoping on behavioral frequency reports. Telescoping is a memory error in the temporal placement of events; that is, an event is remembered, but the remembered date of the event is inaccurate. This uncertainty about the dates of events leads respondents to report events mistakenly as occurring earlier or later than they actually occurred. Bounding reduces telescoping errors in two ways. First, bounding takes advantage of the information collected earlier to eliminate the possibility that respondents report events that occurred outside a given reference period. Second, bounding provides a temporal reference point in respondents' memory, which helps them correctly place an event in relation to that reference point.

A number of specific bounding procedures have been discussed in the survey literature. The bounding interview procedure was first developed by John Neter and Joseph Waksberg in the 1960s in a study of recall of consumer expenditures (they call it "bounded recall"). The general methodology involves completing an initial *unbounded* interview in which respondents are asked to report events that occurred since a given date. In the subsequent *bounded* interviews, the interviewer tells the respondents the events that had been reported during the previous interview and then asks for additional events occurring since then. In other words, the information collected from each bounded interview is compared with information collected during previous interviews to ensure that the earlier reported events are not double counted.

For example, suppose panel respondents are interviewed first in June and then in July. The June interview is unbounded, where respondents are asked to report events that occurred in the previous month. The July interview is bounded. Interviewers would first inform respondents of the data they had provided in June and would then inquire about events that happened since then. Often the data from the initial unbounded interview are not used for estimation but are solely used as a means for reminding respondents in subsequent interviews about the behaviors that have already been reported.

Neter and Waksberg demonstrated in their study that bounding effectively reduced 40% of telescoping on expenditures and 15% on the number of home improvement jobs. This finding encourages panel or longitudinal surveys to employ the bounding technique to reduce the effect of telescoping. The National Crime and Victimization Study (NCVS) is one example. In its redesign, NCVS uses the first of its seven interviews to "bound" the later interviews. There is some evidence suggesting that this bounding technique reduces the likelihood of respondents reporting duplicate victimizations.

The bounding procedure proposed by Neter and Waksberg requires multiple interviews; thus, it is viable only for longitudinal or panel surveys. For one-time surveys, researchers have proposed bounding respondent memory by first asking about an earlier period and then about the more current period. For instance, within a single health interview, respondents are first asked about their health behavior in the previous calendar month and then asked about the same events in the current calendar month. One study shows that bounding within a single interview with two questions reduces reports by between 7% and 20% for health-related behaviors. It reduces telescoping by about 30% to 50% for trivial events, such as purchasing snacks.

Bounding also reduces telescoping error by providing a cognitive reference point in respondents' memory. The initial unbounded interview in Neter and Waksberg's procedure serves a cognitive function for the respondents who recall the last interview and then use that to ascertain whether an event occurred since then. Similarly, the single-interview bounding technique uses the first question to create temporal reference points that assist the respondent in correctly placing an event. A related technique to create a reference point is to use significant dates or landmark events. Landmark events such as New Year's Day, political events, and personally meaningful events (such as a graduation, a wedding, or a local flood) have been used to bound respondents' memory. Research shows that bounding with these landmark events or personally meaningful events significantly reduced incidence of telescoping.

However, bounding with landmark events has its own problems. First, the landmark events might be telescoped forward in one's memory. Second, the landmark events that survey researchers use in a questionnaire might not be equally salient for all respondents interviewed. Thus, subgroup differences might exist in the extent of telescoping error with landmark events, which further distorts comparisons among subpopulations.

Bounding has been shown to be effective in reducing forward telescoping errors and external telescoping errors, but it is less effective in reducing errors resulting from backward telescoping or internal telescoping. In addition, it does not address the effect of forgetting and other types of errors related to retrieving temporal information from long-term memory. Additional research is needed to further investigate the mechanism and the effectiveness of bounding on reducing telescoping error.

Ting Yan

See also Measurement Error; Retrieval; Telescoping

Further Readings

Gaskell, G. D., Wright, G. D., & O'Muircheartaigh, C. A. (2000). Telescoping of landmark events: Implications for survey research. *Pubic Opinion Quarterly*, *64*, 77–89.

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory & Cognition*, *11*, 114–120.

Neter, J., & Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, *59*, 18–55.

Sudman, S., Finn, A., & Lannom, L. (1984). The use of bounded recall procedures in single interviews. *Public Opinion Quarterly*, *48*, 520–524.

Branching

Branching is a questionnaire design technique used in survey research that utilizes skip patterns to ensure that respondents are asked only those questions that apply to them. This technique allows the questionnaire to be tailored to each individual respondent so that respondents with different characteristics, experiences, knowledge, and opinions are routed to applicable questions (e.g., questions about a treatment for diabetes are only asked to respondents who have been diagnosed with diabetes).

Branching also is used to ask respondents to choose among a large number of response options without requiring them to keep all the response options in working memory (e.g., respondents can be asked whether they identify with the Republican or Democratic party and then asked how strongly they identify with the relevant party in follow-up questions).

Branching can be conditional, compound conditional, or unconditional. In conditional branching, a single condition is met where routing occurs based on the answer to a single question (i.e., if the answer to question #1 is "No," then skip to question #3). In compound conditional branching, more than one condition must be met. The branching in this case is dependent on multiple answers, and routing occurs based on a combination of answers (i.e., if the answer to question #1 is "Yes" *or* the answer to question #2 is "Yes," skip to question #5). Unconditional branching is a direct statement with no conditions, often used to bring the respondent back to a specific point in the main survey after following a branching sequence. The approaches to branching differ depending on survey administration.

As a general rule, computer-assisted data collection (i.e., Internet surveys or computer-assisted self, telephone, or personal interviews) allows for more complex branching than paper-and-pencil data collection. Branching can be accomplished in computer-assisted survey instruments using programmed Boolean logic statements (i.e., if (question #) (state condition, such as $=$, $<$, $>$) (value), then (skip to question #)). Branching in paper-and-pencil survey instruments cannot make use of these technological complexities. Rather, it requires the appropriate placement of visual cues to guide respondents or interviewers through the branching instructions. Some common visual layouts include using arrows, placing the branching instructions within approximately nine characters of text (within foveal view), using enlarged, bold, and/or italicized font, and changing the background color. Two additional techniques that can be employed to guide the respondent or interviewers through paper-and-pencil branching instructions are the *prevention technique* and the *detection technique*. In the prevention technique, respondents are educated before reaching the branching instruction by including statements to remind them to look for instructions. In the detection technique, respondents are able to detect any branching errors they may have made through the use of feedback, such as inserting an additional branching instruction before the question that is supposed to be skipped, allowing them to correct the error and follow the instruction as intended.

There are two types of errors associated with branching. *Errors of omission* occur when respondents skip questions that were intended for their completion and result in item nonresponse for those items that were inadvertently skipped. Conversely, *errors of commission* occur when respondents provide answers to questions that were not intended for their completion. Accurate computer-assisted survey programming and proper paper-and-pencil survey visual layout of branching instructions can significantly reduce or even eliminate these errors.

Mindy Anderson-Knott

See also Bipolar Scale; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Errors of Commission; Errors of Omission; Missing Data

Further Readings

Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nichols, W. L., et al. (Eds.). (1998). *Computer assisted survey information collection*. New York: Wiley.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method 2007 update with new Internet, visual, and mixed-mode guide* (2nd ed.). Hoboken, NJ: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2002). *Survey nonresponse*. New York: Wiley.

Bureau of Labor Statistics (BLS)

The Bureau of Labor Statistics (BLS) is an agency within the U.S. Department of Labor (DOL) that is charged with collecting, processing, analyzing, and disseminating essential statistical data about business, finance, employment, and the economy. Other government agencies and many organizations in the private and public sectors heavily rely upon BLS to provide reliable data that is both sweeping in its scope and timely. Its parent organization, the DOL, counts on the BLS to serve as its statistical resource, as does the rest of the federal executive branch, Congress, academic researchers, subnational governmental bodies, private business, labor interests, and ultimately the American public.

BLS has adopted as part of its mission the continual effort to remain relevant to contemporary social and economic issues. It strives for impartiality and data integrity in its statistical reporting. Specifically, BLS follows the Office of Management and Budget's Statistical Policy Directive. Historically, the BLS was established in the late 19th century's period of national expansion and growing economic complexity. The American economy was, and still remains, a rich phenomenon that is accompanied by a large amount of raw data output that can shed light on various aspects of the whole.

In an effort to synthesize the expanse of data into digestible form, BLS conducts survey programs, either themselves or through contracts with the U.S. Bureau of the Census or a cooperating state agency. BLS will then release the gathered data in monthly, quarterly, and annual publications or in periodically published topical reports. Both the chronologically issued reports and the special publications are available in a variety of media including disks and microfiche; however, the most widely used forum for their dissemination is the BLS Web site. Furthermore, the data are available on the Internet at the federal government's multi-agency statistical depository Web site. In addition to these national level reports, the six BLS regional offices (Atlanta, Boston, Chicago, Dallas, Philadelphia, and San Francisco) make available unique data as well. While other government agencies work in the economic data area, notably including the Department of Commerce's Bureau of Economic Analysis and the Federal Reserve Board, it is BLS that offers the most diverse data on the economy. BLS leadership has divided its survey programs into six categories: (1) employment and unemployment, (2) prices and living conditions, (3) compensation and working conditions, (4) productivity and technology, (5) employment projections, and (6) international programs.

Mass media outlets frequently report the work of the BLS on topics that interest a great number of citizens. However, in the process of editing and summarizing the data for the sake of brevity, the media rarely explain the methods by which the information is acquired. The primary survey instrument used by the BLS to gather both employment and unemployment data and compensation and working conditions data is their Current Population Survey (CPS). The CPS is notable because of its sample size and its steady ongoing form, which allows for time series analysis of its results. The survey's 60,000-person sample is drawn from the civilian noninstitutionalized population of the United States that is at least 16 years of age. The basic labor force data are gathered monthly, and special topics are covered on a periodic basis. Because of BLS's compliance with federal privacy guidelines, microdata from individual respondents are not made available. Rather, the data are reported in summary table and aggregate analyses. Information is available for researchers on the population's employment status, broken down by the categories of age, sex, race, Hispanic identity, marital status, family relationship, and Vietnam-era veteran status. The individuals' occupations, industry, class of worker, hours of work, full-time or part-time status, and reasons for working part-time are also included. There are questions posed that are unique to multiple jobholders and discouraged workers as well. The special topic surveys are myriad; they include subjects such as the labor force status of working women with children, and disabled veterans; and also information on work experience, occupational mobility, job tenure, educational attainment, and school enrollment of workers. The results of this survey can be found in BLS-produced sources including the following: *The Employment Situation*, *Employment and Earnings*, *Usual Weekly Earnings of Wage and Salary Workers*, and the *Monthly Labor Review*. Indeed, uses for the data are as diverse, including measuring the potential of the labor supply, determining factors affecting changes in labor force participation of different population groups, and the evaluation of wage rates and earnings trends.

Other than the unemployment rate, perhaps the most widely recognizable output from BLS surveying is that used to calculate the Inflation Rate. The inflation rate is the percentage change in the Consumer Price Index from the preceding year. The BLS collects and processes data on the prices of thousands of goods and services every month, data that in turn produces the cost of a "basket of goods" for a consumer. Additionally, the cost of a "basket of goods" for a firm rather than a consumer is used to calculate the analogous Producer Price Index. Survey work on consumer spending habits, as well as imports and exports, rounds out the BLS's efforts to track prices and living conditions. Notable other statistical output from BLS includes the *Quarterly Labor Productivity Report*, which uses data from the Current Employment Survey, the National Compensation Survey, and the Hours at Work Survey; as well as the *Occupational Outlook Handbook*. The handbook is administered by the Office of Occupational Statistics and Employment Projections and contains information summarizing the working conditions and career prospects of established occupations.

Matthew Beverlin

See also Current Population Survey (CPS)

Further Readings

Fedstats: http://www.fedstats.gov

U.S. Department of Labor, Bureau of Labor Statistics: http://www.bls.gov

Busies

Busies are a survey disposition that is specific to telephone surveys. They occur when the interviewer or a predictive dialer dials a number in the sampling pool and encounters a busy signal. Busies can be considered a positive outcome because they often indicate (a) that the telephone number is in service, and (b) that a person likely can eventually be reached at the number.

Busies can usually be considered a temporary disposition code because the presence of a busy signal is not sufficient to establish whether the respondent or household is eligible for the survey (i.e., busies are cases of unknown eligibility). As a result, it is important to have the interviewer redial the number. One common sample management strategy is to have the number redialed immediately, thus ensuring that the number was dialed correctly and making it possible to reach the person using the phone if he or she was in the process of finishing the call. However, depending on the sample management rules used by the survey organization, busies often also are redialed later in the same interviewing session and on a variety of other days and times in order to maximize the chances of reaching a person. Busies normally are considered a final survey disposition only if a busy signal is the outcome of all call attempts (i.e., the number is always busy) or the only other call outcome is "ring–no answer."

A potential problem in coding busy signals is that they can be confused with *fast busy* signals. These fast busy signals are sometimes used by a number of telephone companies to identify nonworking telephone numbers and can also occur when heavy call volumes fill all of the local telephone circuits. Fast busy case dispositions often are considered final dispositions and ineligible numbers, and thus they usually have a survey disposition code that is different from the code used for normal busies. Telephone interviewers need to understand the difference between busies and fast busy signals, along with the different dispositions of cases that reach busies and fast busy signals. This knowledge will ensure that interviewers code the ineligible, fast busy cases appropriately and will prevent interviewers from making unnecessary additional call attempts on these cases.

Matthew Courser

See also Fast Busy; Final Dispositions; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

C

CALLBACKS

Callbacks are a survey disposition that is specific to telephone surveys. They are a common temporary survey disposition because fewer than half of all completed interviews occur on the first dialing of a case. Callbacks happen for a number of reasons. For example, an interviewer might dial a telephone number in the sampling pool and be told that the designated respondent is not available to complete the interview at the time of the call. In other cases, the interviewer might reach the designated respondent but learn that he or she would prefer to complete the interview at another time. A callback might also occur if an interviewer dials a telephone number and reaches an answering machine or a voicemail service. Callbacks are considered a positive outcome because they usually indicate that the household or designated respondent is eligible and that an interview is likely to be completed with the respondent if the interviewer is able to reach him or her at a good time. Cases coded with the callback disposition usually are considered eligible cases in calculating survey response rates because the interviewer has been able to determine that the household or designated respondent meets the qualifications set by the survey researcher for completing the interview.

Callbacks can occur for multiple reasons, and as a result the callback disposition often is further categorized into a *general* callback disposition and a *specific* callback disposition. In a general callback, the interviewer learns that the designated respondent is not available at the time of the call but does not learn anything that would help him or her determine the best time to reach the designated respondent. In other cases coded with a general callback disposition, the interviewer may obtain some information about when to next make a call attempt on the case (such as "evenings only" or "before 2:30 p.m.") but is not able to make an appointment to contact the designated respondent at a definite day or time. In a specific callback, however, the interviewer learns enough to set a definite day and time for the next call attempt (such as, "appointment set for 2:30 p.m. tomorrow"). Aside from learning the day and time for subsequent call attempts, interviewers also should attempt to obtain other information that might increase the chances of converting the callback into a completed interview. This information might include the name and/or gender of the designated respondent, or any other information that might help the interviewer reach the designated respondent on subsequent call attempts.

Because cases coded with the callback disposition are eligible and continue to be processed in the sampling pool, information learned during previous call attempts about when to contact the designated respondent can be used to better target subsequent call attempts by the interviewer. For a specific callback, additional call attempts should occur at the appointment time set by the respondent; additional call attempts on a general callback in which little is known might be made at a variety of other days and times in order to increase the chances of reaching the designated respondent and/or to learn more about how to target additional call attempts.

Matthew Courser

See also Busies; Calling Rules; Designated Respondent; Final Dispositions; Noncontacts; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Caller ID

Caller ID is a telephone service in the United States that transmits the caller's name and/or telephone number to the called party's telephone. Today most telephones come with caller ID capabilities, and telephone companies regularly offer the service for little or no cost as part of their monthly service packages. Caller ID consists of two elements: the calling number and the subscriber name. This information appears on a person's telephone or display unit. Caller ID service lets you identify yourself to the person you are calling and lets you see who is calling before you answer the phone. It is estimated that more than half of all households in the United States have caller ID. Because this technology allows people to see who is calling, it is frequently used to screen unwanted calls, including those from survey research organizations. More and more people are using caller ID technology and caller ID–based services to screen incoming calls. A variety of call screening services or devices allow households to selectively or arbitrarily reject anonymous callers or any phone number that is not preidentified to ring through.

The Federal Communications Commission (FCC) has developed national caller ID rules. These rules allow subscribers to block or prevent their names and numbers from being displayed permanently or on a call-by-call basis. Conversely, the FCC rules require telemarketers to transmit caller ID information and prohibit them from blocking such information. Calls to emergency lines, such as 911, are exempt from federal caller ID rules and are governed by state rules and policies.

Caller ID technology and related call-blocking services will certainly continue to grow in popularity. Therefore researchers must continue to analyze the impact of this technology on response rates and to experiment with using caller ID technology to improve response rates. Although research firms are not required to send caller ID information, there is some experimental evidence that response rates may be improved by sending the survey firm name or an 800-number as their caller ID tag.

Linda Piekarski

See also Call Screening; Federal Communications Commission (FCC) Regulations; Noncontacts; Privacy Manager

Further Readings

Link, M., & Oldendick, R. (1999). Call screening. *Public Opinion Quarterly*, *63*, 577–589.

Trussell, N., & Lavrakas, P. J. (2005, May). *Testing the impact of caller ID technology on response rates in a mixed mode survey*. Paper presented at 2005 American Association for Public Opinion Conference, Miami Beach, FL.

Call Forwarding

Call forwarding is a feature on most U.S. and international telephone networks that allows an incoming call to be redirected to one or more other telephone numbers as directed by the subscriber. This feature is popular with individuals who want or need to be reached when they are not at home or want to avoid the delays inherent with answering machines and voicemail. The use of call forwarding features can cause problems for telephone survey researchers. When an incoming call has been forwarded to another location, the called party may be less willing to participate in a survey at that location. When a call is forwarded to a cell phone in the United States, the called party will incur a cost in terms of dollars or minutes and may be in a location or other circumstance that is incompatible with survey participation.

Standard call forwarding transfers all calls from phone number A to phone number B. Special types of call forwarding are also available. Call forwarding can automatically route calls that are not answered within a designated number of rings or when the line is busy to another telephone number. Finally, call forwarding can transfer only those calls coming from a select set of telephone numbers. Remote access to call forwarding allows customers to activate or deactivate call forwarding from any telephone equipped with touch tone. In the North American Numbering Plan, vertical service codes, such as * 72 for activation, are used to control call forwarding. Usually, the forwarded line rings once, to remind anyone there that calls are being redirected.

The fee structures associated with placing a call to a called party who has his or her number forwarded can be subtle. For example, in the United States, Person A in Pittsburgh calls Person B in Chicago, who has forwarded his calls to Person C in Los Angeles. Person A will be charged for a long-distance call from Pittsburgh to Chicago, and Person B will be charged for a long-distance call from Chicago to Los Angeles. Call forwarding from a landline number to a cell phone will result in additional costs to respondents and problems associated with location of the respondent at the time of contact. These charges and unexpected circumstances may make respondents less likely to cooperate in a survey when reached at a telephone number or location other than their residences. Since sample suppliers routinely remove numbers assigned to wireless services from their databases, most of the cell phones encountered in telephone surveys are likely the result of call forwarding. Researchers should attempt to identify these cell phones early in the interview process and offer alternative means for completing the interview.

Finally, call forwarding may mean that an interview is completed in a location other than that associated with the telephone number dialed. For example, in the case of the areas affected by the hurricanes of 2005, call forwarding was included in the list of waived services that customers of BellSouth could consider using during their displacement. Also, a telephone company sometimes briefly uses call forwarding to reroute calls from an old number to a new number after a customer moves or ports his or her number to a new provider.

A problem caused by call forwarding that researchers doing surveys of the general population must address occurs when the original number dialed is a business number and it is forwarded to a residential number. In these cases, the household that actually is reached is not considered eligible because it was reached by sampling a nonresidential number. To determine when this happens, interviewers need to verify with the respondent that she or he has been reached at the number that was dialed.

Linda Piekarski

See also Federal Communications Commission (FCC) Regulations; Number Portability; Telephone Consumer Protection Act of 1991

CALLING RULES

Telephone survey researchers often utilize a set of guidelines (or *calling rules*) that dictate how and when a sample unit should be contacted during the survey's field period. These rules are created to manage the sample with the goal of introducing the appropriate sample elements at a time when an interviewer is most likely to contact a sample member and successfully complete an interview. In telephone surveys, calling rules are typically customized to the particular survey organization and to the particular survey and should be crafted and deployed with the survey budget in mind.

Calling rules are a primary mechanism that researchers can use to affect a survey's response rate. All else equal, making more dialing attempts will lower noncontact-related nonresponse, thereby yielding a higher response rate. In general, the more call attempts placed to a telephone number, the more likely someone will eventually answer the phone, thereby giving the survey organization's interviewers the opportunity to try to complete an interview. However, the trade-off to making more and more phone calls is the additional costs incurred with each call, both in terms of interviewers' labor and the toll charges related to the calls.

Since all surveys have finite budgets and resources that must be allocated for dialing attempts, resources allocated for these purposes cannot be used for other important purposes, such as additional questionnaire testing or development or gathering data from larger sample sizes. This competition for survey resources, along with the tension between achieving higher response rates with more calls made and the added expenditure of these additional call attempts illustrates the importance of a well-thought-out approach to the development and implementation of calling rules to manage a telephone survey sample.

When examining calling rules, an important distinction is often made between first call attempts to a sample member, or *cold calls*, versus subsequent calls to sample members, or *callbacks*. The importance of this distinction lies in the different information that is available to the survey researcher to establish calling rules.

In the case of first call attempts, little information exists about the sample member, including no information about the effectiveness of calls previously placed to that sample member. For subsequent call attempts, however, the call history for the sample numbers can be utilized to refine the placement of calls to these sample members. Consequently, calling rules for subsequent calls often differ from the calling rules used to place initial calls.

These calling rules, regardless of whether they apply to first call attempts or subsequent call attempts, can be classified into two different types: *ranked category type* calling rules and *priority scoring type* calling rules. Each type denotes an inherent property of calling rules, which is to create some calling order for survey administrators to follow with active samples.

Ranked Category

In the case of ranked category calling rules, the sample is categorized into independent (nonoverlapping) cohorts, based on sample member characteristics and/or previous call outcomes, and then ranked in order of the most likely categories to lead to a contacted sample member. For example, a simple ranked category calling rules system might suggest that previously reached sample members, answering machines, and ring–no answers are categorized as such and then should be called in that order. More complicated ranked category systems would classify the sample into more specialized categories and, therefore, have more elaborate calling rules to process the sample. As an example, for sample members who have yet to be contacted, categories could be created that take into account the time and day that previous calls had been made. Calling rules could then dictate that future calls should be made at times and days on which previous calls had not been attempted. Once a call attempt is made under a ranked category calling rules system, assuming that the sample member remains part of the active sample, the information gained from the last call is incorporated into the information set for that sample member. This additional information collected from the last call is used to recategorize the sample member, possibly into a different sample category.

Ranked category calling rules can be implemented using computer-assisted telephone interviewing (CATI), but they can also be implemented without the use of computers, making them an effective means by which to control and process the sample. However, a drawback to the use of ranked category calling rules is the multitude of different categories that may be necessitated and then the elaborate system of calling rules that would be developed to rank these categories.

Priority Scoring

Priority scoring calling rules differ from ranked category calling rules in that, with priority scoring, it is not necessary to categorize the sample into discrete, nonoverlapping categories. Instead, the information collected for each sample member is used in a multivariate model, typically a logistic regression model, to estimate the probability of the next call attempt leading to a contact and/or completion, conditioned on relevant information. Using the estimated coefficients from this multivariate model, the probability of contact or completion can be calculated for any possible permutation of the conditioning information set. These probabilities are then used to order the sample, from the highest probability calls to the lowest, with the highest probability calls being made first.

For example, a sample member who has been called three times previously, once in the afternoon and twice in the evening, with the outcomes of one ring–no answer, one busy signal, and one callback may have a contact probability of 0.55 if the next call attempt is placed in the evening. Another sample member who has been called five times previously, once in the morning, twice in the afternoon, and twice in the evening, with the outcomes of three ring–no answers, one busy signal, and one callback may have a contact probability of 0.43 if the next call attempt is placed in the evening. Although both contact probabilities indicate a fairly high likelihood of reaching these sample members in the evening, the contact probability for the first sample member is higher, so that priority scoring calling rules would rank that sample member higher in the calling queue.

Once the call attempt is made, assuming that the sample member continues to be part of the active sample, the information gained from this call attempt updates the sample member's information set. This updated information is used to calculate an updated contact probability, which is then used to rank order the sample member in the existing active sample.

Priority scoring calling rules are a model-based approach that, once implemented, can effectively manage samples, continually updating contact probabilities to deliver the most likely sample members to be contacted. Moreover, not only can the conditioning information be used to determine jointly the effects of that information on contact probabilities, but also, to the extent there are interaction effects with the conditioning information, these effects can be explicitly modeled with a priority scoring system of calling rules. However, a drawback to the use of priority scoring is the requirement of CATI, both because the multivariate model that serves as the basis for the priority scoring calling rules typically is a function with numerous covariates and also because the calculation and updating of contact probabilities does not lend itself to manual calculation.

Conditioning Information

In order to develop ranked category calling rules or priority scoring calling rules, some prior understanding of the likelihood of contacting sample members, given the condition information, must be available. Typical conditioning information that is used can be classified into external information about sample members—for example, demographics, telephone number or exchange information—and call history information about sample members. Call history information that has been used for initial calls includes the time of day and the day of the week the first call is made. Call history information that has been used for subsequent calls includes not only the information used for first calls but also the number of previous calls that have been made, the length of time between the last call and the next call, the disposition of the previous call, the entire history of call dispositions, and the time and days that previous calls were made.

Typically, previous survey experience governs not only the use of conditioning information either to categorize or to score the sample, but also how this conditioning information impacts the contact probabilities. To the extent that the population for a survey has been studied before, the use of the conditioning information from that prior survey can be used to develop calling rules for subsequent surveys of that same population. However, to the extent the survey researcher is studying a population for the first time, the only avenue open for the development of calling rules may be to base them on a survey of a population that is similar, albeit unrelated.

Jeffery A. Stec

See also Callbacks; Cold Call; Computer-Assisted Telephone Interviewing (CATI); Contacts; Elements; Field Period; Sample Management; Telephone Surveys

Further Readings

Harpuder, B. E., & Stec, J. A. (1999). Achieving an optimum number of callback attempts: Cost-savings vs. non-response error due to non-contacts in RDD surveys. *Proceedings of the Section on Survey Research Methods* (pp. 913–918). Alexandria, VA: American Statistical Association.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Massey, J. T., Wolter, C., Wan, S. C., & Liu, K. (1996). Optimum calling patterns for random digit dialed telephone surveys. *Proceedings of the Section on Survey Research Methods* (pp. 485–490). Alexandria, VA: American Statistical Association.

Reedman, L., & Robinson, M. (1997). An improved call-scheduling method using call history and frame information. *Proceedings of the Section on Survey Research Methods* (pp. 730–735). Alexandria, VA: American Statistical Association.

Stec, J. A., Lavrakas, P. J., Shuttles, C. D., Daily, G., Yancey, T., & Watkins, R. (2007, May). *Gaining efficiencies in scheduling callbacks in large RDD national surveys*. Paper presented at the 2007 American Association for Public Opinion Research conference, Anaheim, CA.

CALL-IN POLLS

A call-in poll is an unscientific attempt to measure public preferences by having radio or television audience members or newspaper readers call a telephone number and register their opinions. Usually a single question is posed, and people are asked to call one phone number in support of a viewpoint and another number in opposition. Call-in polls are used by some media organizations as a way to measure public opinion and get the audience involved. But they are very problematic from a data quality standpoint and should not be referred to as "polls."

A major problem with call-in polls is that the participants are entirely self-selected. Only those people who tuned in to that particular broadcast at that time, or read that newspaper, can be included. Further, those who make the effort to participate are often very different from those who do not. That is because participants are usually more interested in the topic or feel very strongly about it. For these reasons, survey researcher Norman Bradburn of the University of Chicago coined the term *SLOP*, which stands for "self-selected listener opinion poll," to refer to call-in polls.

Another big problem is that call-in polls are open to manipulation by any individual or group with a vested interest in the topic. With no limit on the number of calls that can be placed, people can call multiple times and groups can set up more elaborate operations to flood the phone lines with calls in support of their point of view. As a result, call-in polls often produce biased results, and their "findings" should be ignored. Legitimate survey researchers avoid the types of bias inherent in call-in polls by selecting respondents using probability sampling techniques.

There are many examples of call-in polls producing distorted results. In one famous example, *USA Today* conducted a call-in poll in 1990 asking its readers whether Donald Trump symbolizes what is wrong with the United States or symbolizes what makes the United States great. *USA Today* reported overwhelming support for Trump, with 81% of calls saying he symbolizes what makes the United States great. Later, *USA Today* investigated the results and found that 72% of the 7,802 calls came from a company owned by a Trump admirer.

Another example comes from a 1992 CBS television program called *America on the Line*, where viewers were asked to call in and register their opinions after President George H. W. Bush's State of the Union address. The views of the approximately 317,000 calls that were tallied were much more pessimistic about the economy than what was measured in a traditional scientific poll conducted by CBS News at the same time. For example, 53% of those who called in to the program said their personal financial situation was worse than 4 years ago, compared with 32% in the scientific poll. The views of those who called in were quite different than those of the general public on a number of measures.

Although those with survey research training know that call-in polls should not be taken seriously, many members of the public do not make a distinction between these pseudo-polls and the real thing. In fact pseudo-polls may be incorrectly seen as even more credible than real polls because they often have much larger sample sizes.

Daniel M. Merkle

See also 800 Poll; Log-In Polls; 900 Poll; Pseudo-Polls; Probability Sample; Self-Selected Listener Opinion Poll (SLOP); Self-Selected Sample; Self-Selection Bias

Call Screening

Call screening is a practice in which many people engage whereby they listen to an incoming message on their answering machine or look on their caller ID to see who is calling before deciding whether or not to answer the call. This behavior is thought to negatively affect survey response rates. Over time, respondents have become increasingly unwilling to participate in surveys or even answer unsolicited telephone calls. This desire for privacy has resulted in legislation such as do-not-call lists and the use of a variety of technological barriers such as answering machines, caller ID, and call blocking to screen incoming calls. These screening devices allow individuals to determine which calls they will answer, making it more difficult for researchers to contact them. Further, individuals who always screen may also be more likely to refuse to participate if and when they are contacted.

More than two thirds of U.S. households have answering machines, and about 18% report always using their answering machine to screen calls. Telephone companies improved on the answering machine as a screening device with the development of caller ID technology. This service displays the caller's name and/or telephone number on a person's phone or caller ID device. It is estimated that more than half of all U.S. households now have caller ID and that nearly 30% always use it to screen calls. Call-blocking services that allow subscribers simply to reject certain numbers or classes of numbers are also growing in popularity.

Owners of these devices and those who regularly use them to screen calls have been shown to be demographically different from the general population. It is not always easy to identify a screening household, particularly if the dialing always results in a noncontact.

A number of approaches are being used by researchers in an attempt to improve contact with screening households. The most common approaches include mailing advance letters (when a phone number can be matched to an address), leaving a message on the answering machine, or transmitting the name of the research firm along with an 800 call-in number. However, when it comes to actually improving contact with these households, the results remain mixed.

Linda Piekarski

See also Advance Letter; Answering Machine Messages; Caller ID; Do-Not-Call (DNC) Registries; Privacy Manager

Further Readings

Link, M. W., & Oldendick, R. W. (1999). Call screening. *Public Opinion Quarterly*, *63*, 577–589.

Tuckel, P., & O'Neill, H. W. (1996). Screened out. *Marketing Research*, *8*(1), 34–43.

Call Sheet

A call sheet is a record-keeping form that is used by telephone survey interviewers to keep track of information related to the calls they make to reach survey respondents. As paper-and-pencil interviewing (PAPI) was replaced by computer-assisted telephone interviewing (CATI), these call sheets moved from being printed on paper to being displayed on the interviewer's computer monitor. The fact that they are named "call sheets" refers to the days when thousands of such call sheets (each one was a piece of paper) were used to control sampling for a single telephone survey.

The information that is recorded on a call sheet—also called "paradata"—captures the history of the various call attempts that are made to a sampled telephone number. Typically these forms are laid out in matrix format, with the rows being the call attempts and the columns being the information recorded about each call. For each call attempt, the information includes (a) the date; (b) the time of day; (c) the outcome of the call (disposition), for example, ring–no answer, busy, disconnected, completed interview, and so on; and (d) any notes the interviewer may write about the call attempt that would help a subsequent interviewer and/or a supervisor who is controlling the sample, for example, "The respondent is named Virginia and she is only home during daytime hours." Since most telephone interviews are not completed on the first calling attempt, the information that interviewers record about what occurred on previous call attempts is invaluable to help process the sample further and effectively.

It is through the use of the call outcome information recorded on the call sheet—and described in detail in the American Association for Public Opinion Research's *Standard Definitions*—that the sample is managed. In the days when PAPI surveys were routinely conducted and the call sheets were printed on paper, supervisory personnel had to sort the call sheets manually in real time while interviewing was ongoing. When a questionnaire was completed, the interviewer manually stapled the call sheet to the top of the questionnaire and then the supervisor removed that case from further data collection attempts. For call sheets that did not lead to completed interviews but also did not reach another final disposition (e.g., disconnected or place of business), the supervisor followed a priori "calling rules" to decide when next to recycle a call sheet for an interviewer to try dialing it again. With the shift to CATI and computer control of the sampling pool (i.e., the set of numbers being dialed) all this processing of the information recorded on call sheets has been computerized. The CATI software serves up the call sheet on the interviewer's monitor at the end of the call for pertinent information to be entered. That information drives other logic in the CATI software that determines whether, and when, to serve up the telephone number next to an interviewer. The information captured on the call sheet is used for many other purposes after the survey ends, including helping to create interviewer performance metrics and calculating survey response rates.

Paul J. Lavrakas

See also Callbacks; Calling Rules; Computer-Assisted Telephone Interviewing (CATI); Interviewer Productivity; Paper-and-Pencil Interviewing (PAPI); Paradata; Response Rates; Sampling Pool; Standard Definitions; Telephone Surveys

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Hansen, S. E. (2008). CATI sample management. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 340–358). New York: Wiley.
Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Capture–Recapture Sampling

Capture–recapture sampling (also referred to as "capture–mark–recapture sampling" or "mark–release–recapture sampling") is a method used to estimate the unknown size of a population. In practice, it is often not feasible to manually count every individual element in a population because of time, budget, or other constraints. And, in many situations, capture–recapture sampling can produce a statistically valid estimate of a population size in a more efficient and timely manner than a census.

The most basic application of capture–recapture sampling consists of two stages. The first stage involves drawing (or capturing) a random sample of elements from a population of unknown size, for example, fish in a pond. The sampled elements are then marked, or tagged, and released back into the population. The second stage consists of drawing another random sample of elements from the same population. The second-stage sample must be obtained without dependence on the first-stage sample. Information from both stages is used to obtain an estimate of the population total.

The capture–recapture technique assumes that the ratio of the total number of population elements to the total number of marked elements is equal, in expectation, to the ratio of the number of second-stage sample elements to the number of marked elements in the sample. This relationship can be expressed as follows:

$$N/C = n/R, \quad (1)$$

where N is the unknown population total of interest, n is the number of elements in the second-stage sample (both marked and unmarked), C is the total number of marked elements from the first-stage sample (i.e., the captures), and R is the number of marked elements found in the second-stage sample (i.e., the recaptures). By solving for N, it is then possible to obtain an estimate of the population total:

$$N = nC/R. \quad (2)$$

Example

A classic example comes from the field of ecology. Suppose the goal is to estimate the size of a fish population in a pond. A first-stage sample of 20 fish is drawn, tagged, and released back into the pond. A second-stage sample of 30 fish is subsequently drawn. Tags are found on 12 of the 30 sampled fish, indicating that 12 fish captured in the first sample were recaptured in the second sample. This information can be used to assign actual quantities to the variables of interest in Equation 1, where $n = 30$, $C = 20$, and $R = 12$. Solving for N using Equation 2 yields the following estimate of the population total:

$$N = nC/R = ((30)(20))/12 = 50.$$

Therefore, the estimated size of the pond's fish population is 50. A more stable estimate of the population total, subject to less sampling variability, can be obtained if multiple second-stage samples are drawn, and estimated totals, computed from each sample, are averaged together.

Assumptions

In order for the capture–recapture sampling technique to produce a valid estimate of a population size, three assumptions must hold:

1. Every population element has an equal probability of being selected (or captured) into both samples.
2. The ratio between marked and unmarked population elements remains unchanged during the time interval between samples.
3. Marked elements can be successfully matched from first-stage sample to second-stage sample.

Assumption 1 holds if simple random sampling is used to capture elements into both samples. A possible violation of this assumption occurs if those who were captured in the first-stage sample have a higher probability of being captured in the second-stage sample, which would lead to overestimation of the population total. Assumption 2 follows from the relationship described in Equation 1. In general, this assumption holds if there is no change in the population, or if the population is closed during the study. However, births or deaths and immigration or emigration are permitted as long as the ratio is preserved.

frame for those people or households in the overlap; instead, they make use of inclusion probabilities that are frame specific (i.e., either CPN frame or LLN frame). Adjustments to the weights for multiple cell phones are made for subscribers in the cell phone sample; similarly, weight adjustments are applied for multiple landlines for households selected from the landline frame. Using the frame-specific adjusted weights, estimates for the variables of interest are derived from the CPO and C&L and the LLO and C&L pieces from the cell phone and landline samples, respectively. The two estimates of the overlap (C&L) are combined via a composite estimator, with the weights chosen to minimize the variance of the statistic of interest.

A simpler but related alternative that avoids having to weight the sample for inclusion in both frames and seems to be used frequently in current practice involves conducting a random-digit dial (or other common sampling technique, such as list-assisted, etc.) of landline numbers. This sample is then augmented by a sample of cell phone numbers that has been screened for cell phone only households. The phone ownership distribution of the combined sample is then weighted using some type of post-stratification weighting technique (such as raking, etc.) to the distribution obtained via a personal interview survey such as the National Health Interview Survey or the Current Population Survey. However, these data are only available at the U.S. national level.

The adjustments and estimators discussed thus far assume complete response, which is not likely in practice. Additional adjustments for nonresponse will be needed in the weights. Of course, it always helps to attempt to reduce nonresponse. Some details of the cell phone numbering systems and plan attributes may be helpful for designing more efficient data collection measures for units included in cell phone samples.

Cell Phone Numbering Systems

Numbering systems or agencies such as the North American Numbering Plan Administration (NANPA) assign banks of numbers to cell phone providers. One main difference in the CPNs between countries is the level of geographic specificity that can be inferred. In some countries (e.g., United Kingdom, Italy), CPNs are organized in two parts: the prefix indicates the cell phone provider and the suffix is the number assigned by that provider to the final user. In the United States, CPNs are organized into three parts: area code, prefix, and suffix. The area code is three digits and indicates specific geographic regions that usually do not cross state boundaries. Generally, there is a strong concordance between place of residence and area code, but because cell phones are portable and national networks exist for many providers, it is possible that the degree of specificity could be limited to the location in which the cell phone contract was initiated. The three-digit prefix generally indicates the cell phone provider and, to a lesser degree, a geographic area within the region of the area code. The four-digit suffix is assigned by the cell phone provider. The assignment rules for these numbers are more ambiguous when compared to that of landlines. In fact, an informal survey of major U.S. providers in 2005 did not reveal any trends or clustering patterns by which CPNs are assigned to new subscribers. However, in many cases company representatives indicated that number assignments are highly proprietary, especially in an era when NANPA is imposing new regulations on number bank allocations based on usage capacity quotas: some prefixes now include suffixes that are either LLNs or CPNs assigned by the same provider (i.e., mixed-use bank) or LPNs or CPNs assigned by different providers (i.e., mixed-provider bank). This ambiguity in number assignment makes methods like the Mitofsky-Waksberg method of limited utility for cell phone samples. Also, unlike LLNs, CPNs are not usually publicly available in phone directories, so list-assisted approaches are also limited for cell phone samples. There are exchange-type codes available within the telephone industry and from vendors who supply samples of cell and landline numbers that can be used by researchers to help determine which of the 1,000-banks contain both cell and landline numbers. There are companies in the United States that now provide samples of cell phone numbers from a frame of 10,000-banks that have already been screened for mixed use.

Cell Phone Services

Cell phone services are generally organized differently, tend to vary more, and change more rapidly than landline phone services. Subscribers access cell phone service through a wide array of contracts and service plans. These contracts can be classified into two broad categories: pre-paid and post-paid. For the pre-paid contracts, limits vary by provider for the amount of time the associated phone number can be retained for accounts that have become dormant (i.e., have not been

Assumption 3 holds if there is no loss of tags and no erroneous matching.

Typically, these assumptions cannot be tested using a two-stage sampling approach. More advanced capture–recapture methods exist that allow these assumptions to be tested, and in some cases, permit certain assumptions to be relaxed. For example, methods have been proposed that consider situations where elements have different probabilities of being captured—a violation of Assumption 1.

1990 Post-Enumeration Survey

One of the most notable applications of capture–recapture sampling occurred during the 1990 Post-Enumeration Survey (PES). The goal of the PES was to evaluate the accuracy of the 1990 Census enumeration. A capture–recapture approach was used to estimate the total number of individuals who were omitted from the census enumeration. The first-stage sample consisted of all individuals who were enumerated in the 1990 Census. Census Bureau records were used to help identify those who were included in the enumeration. In the second stage, an area probability sample of household blocks was drawn. Individuals within sampled households were interviewed, and census records were checked to determine whether or not they had been included in the census. By counting the number of individuals in the second-stage sample who were left out of the census enumeration, an estimate of the total census undercount was obtained.

Other applications of capture–recapture sampling have been applied to estimating birth and death rates, estimating the number of HIV-infected drug users, estimating the incidence of stroke, and estimating salmon spawning escapement.

Joseph W. Sakshaug

See also Bias; Census; Elements; Sampling; Simple Random Sample

Further Readings

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*(4), 783–791.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, *88*, 1047–1060.

Le Cren, E. D. (1965). A note on the history of mark-recapture population estimates. *Journal of Animal Ecology*, *34*(2), 453–454.

Shapiro, S. (1949). Estimating birth registration completeness. *Journal of the American Statistical Association*, *45*, 261–264.

CASE

The term *case* refers to one specific element in the population of interest that has been sampled for a survey. A "completed" case contains the responses that were provided by that respondent for the questionnaire used in that survey. A case may be an individual, a household, or an organization. Being able to identify each individual respondent can be critical for the conduct of the survey. Assignment of a unique *case number* identifier associated with each individual sampled element should be done in every survey. Although most computer-assisted surveys assign a *respondent number*, it should not be confused with assignment of a case number. As a general rule, case numbers are assigned before a questionnaire is distributed, while respondent numbers are assigned when a respondent is contacted and an attempt is made to complete the survey.

Prior to data collection, a simple case number may be assigned sequentially to each questionnaire before being distributed for completion. The case number can also be used to identify any number of background characteristics of the individual or household to which the survey was distributed—such as census block, zip code, or apartment or single-family home. Assignment of a case number should not be used to compromise the confidentiality of either those who complete the survey or the information they provide. During data processing, the case number can be used to assist in coding open-ended responses and conducting edit checks on the data set, such as verifying information that is outside the normal response range or that is inconsistent with other data in the case record. In those designs for which respondents may be contacted at a future date, the unique case number can be used to ensure that responses to future surveys are linked to the correct respondent.

Dennis Lambries

See also Coding; Completion Rate; Element; Respondent

Case-Control Study

Case-control studies measure the association between the exposure to particular risk factors and the occurrence of a specific disease. These types of studies are common in public health and medical research. The basic premise of such studies is the comparison of two groups: "cases," individuals who have a particular disease of interest to the researcher, and "controls," who do not have the disease.

In case-control studies, individuals in the case group are selected and matched to persons in the control group on a common set of characteristics that are not considered to be risk factors for the disease being studied. These characteristics are frequently demographic variables such as age, gender, education, income, and area of residence. Comparisons across the case-control pairs are made, examining hypothesized risk factors for a particular disease. For example a case-control study of heart disease among women may compare cases and controls on their level of exposure to factors thought to influence the risk of heart disease such as family history of heart disease, smoking, cholesterol, high blood pressure, diet, and exercise. These differences are usually assessed using statistical tests.

Data for case-control studies is typically collected by interviewing or surveying the cases and the controls. Individuals in both groups are asked the same series of questions regarding their medical history and exposure to factors that are considered to increase the risk of developing the disease in question. Data may also be collected from medical records.

The advantages of case-control studies include the following:

- Data collection does not typically require medical tests or other intrusive methods.
- The studies are typically inexpensive to conduct in comparison to other methods of data collection.
- They are good for examining rare diseases because the investigator must identify cases at the start of the research rather than waiting for the disease to develop.
- Case-control studies allow for the examination of several risk factors for a particular disease at the same time.

As with all research studies, there are some significant disadvantages as well, including the following:

- Data on exposure and past history is subject to the individual's memory of events.
- It can be difficult to confirm and/or measure the amount of exposure to a particular risk factor of interest.
- Defining an appropriate control group can be difficult, especially if the risk factors for a particular disease are not well defined.
- Case-control studies are not good for diseases that result from very rare risk factors (rare exposures) unless there is a high correlation between the disease and the exposure.

Katherine A. Draughon

See also Case; Control Group; Research Design

Further Readings

Hennekens, C. H., Buring, J. E., & Mayrent, S. L. (Eds.). (1987). *Epidemiology in medicine.* Boston: Little, Brown.

Cell Phone Only Household

The widespread availability of cell phone service and the relatively low cost of such service means that some people are now indifferent as to whether they make a call on a landline or a mobile telephone. In fact, many people have substituted one or more wireless cell phones for their traditional household wired telephones (also called "residential landline telephones"). These cell phone only households pose a problem for most major survey research organizations in the United States because cell phone numbers are not typically included when conducting random-digit dial (RDD) telephone surveys in the United States.

The Telephone Consumer Protection Act of 1991 prohibits the use of autodialers in the United States when calling cell phones; therefore, the inclusion of such telephone numbers would be very expensive for most survey call centers because of the requirement to have interviewers dial the cell phone numbers manually. In addition, nonresponse rates may be high because most cell phone owners do not expect to receive survey calls on their cell phones, and some cell phone owners must pay to receive calls.

The inability to reach cell phone only households has potential implications for coverage bias in random-digit dialed telephone surveys. Coverage bias may exist if cell phone only households are not included in survey sampling frames and if persons living in cell phone only households differ on the survey variables of interest from persons living in households with landline telephones.

The National Health Interview Survey (NHIS) provides the most up-to-date estimates regularly available from the U.S. federal government concerning the prevalence and characteristics of cell phone only households. This cross-sectional, in-person, household survey of the U.S. civilian noninstitutionalized population, conducted annually by the National Center for Health Statistics of the Centers for Disease Control and Prevention, is designed to collect information on health status, health-related behaviors, and health care utilization. However, the survey also includes information about household telephones and whether anyone in the household has a working cell phone. Approximately 40,000 household interviews are completed each year.

NHIS data permit an analysis of trends in the prevalence of cell phone only households in the United States since 2003. The percentage of cell phone only households doubled from 2003 to 2005, and as of 2006, approximately 11% of U.S. households were cell phone only. The rate of growth in the size of this population has not slowed, increasing at a compound growth rate of more than 20% every 6 months. Cell phone only households now compose the vast majority of non-landline households. More than 80% of non-landline households have cell phone service in the household, and this proportion also continues to increase; the proportion was 62% during the first 6 months of 2003. This largely reflects the fact that the percentage of households without any telephone service has remained unchanged, whereas the percentage of cell phone only households has increased.

Since the NHIS began collecting data on cell phone only households and the persons who live in such households, the prevalence of cell phone only adults has been greatest for adults 18–24 years of age, adults renting their homes, and adults going to school. Men are more likely than women to be living in cell phone only households. Hispanic adults are slightly more likely to be living in cell phone only households than are non-Hispanic white adults or non-Hispanic black adults. Adults living in the Midwest, South, or West are more likely to be living in cell phone only households than are adults living in the Northeast. Adults living in urban households are more likely than adults living in rural households to be in cell phone only households.

Adults working at a job or business in the week prior to the interview are also more likely to live in cell phone only households than adults who are keeping house or are unemployed or doing something else. Yet, adults living in poverty are more likely than higher income adults to be living in cell phone only households.

Adults living with unrelated roommates are more likely to live in cell phone only households than adults with other living arrangements. Looking at other family structure subgroups, adults living alone are more likely to be cell phone only than are adults living with other related adults or adults living with children.

Despite the differences in demographic characteristics between persons living in households with landline telephones and persons living in cell phone only households, the potential for coverage bias in population-based surveys of adults has been found to be small so far. Estimates from health surveys and from political polls that did not include data from the cell phone only population have not been substantially biased when proper survey weighting and estimation strategies have been employed. However, as the size of the cell phone only population grows in this rapidly changing technological environment, the potential for coverage bias may also increase.

If this occurs, survey researchers will need to determine how best to add cell phone only households to their sampling frames. This may occur by calling cell phones directly or by conducting multi-mode surveys that reach cell phone only households by mail, Internet, and/or in person. Methodologies are being developed currently for conducting surveys on cell phones and for combining sampling frames that use multiple modes.

Stephen J. Blumberg

See also Cell Phone Sampling; Coverage Error; National Health Interview Survey (NHIS); Telephone Consumer Protection Act of 1991; Telephone Households

Further Readings

Blumberg, S. J., & Luke, J. V. (2007). Coverage bias in traditional telephone surveys of low-income and young adults. *Public Opinion Quarterly*, *71*(5), 734–749.

Blumberg, S. J., Luke, J. V., & Cynamon, M. L. (2006). Telephone coverage and health survey estimates: Evaluating the need for concern about wireless substitution. *American Journal of Public Health*, *96*, 926–931.

Blumberg, S. J., Luke, J. V., Cynamon, M. L., & Frankel, M. R. (2008). Recent trends in household telephone coverage in the United States. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 56–86). New York: Wiley

Ehlen, J., & Ehlen, P. (2007). Cellular-only substitution in the United States as lifestyle adoption: Implications for telephone survey coverage. *Public Opinion Quarterly*, *71*(5), 717–733.

Keeter, S. (2006). The impact of cell phone noncoverage bias on polling in the 2004 presidential election. *Public Opinion Quarterly*, *70*, 88–98.

Keeter, S., Kennedy, C., Clark, A., Tompson, T., & Mokrzycki, M. (2007). What's missing from national landline RDD surveys? The impact of the growing cell-only population. *Public Opinion Quarterly*, *71*(5), 772–792.

Nielsen Media Research. (2005). *Cell Phone Sampling Summit II.* Retrieved March 24, 2008, from http://www.nielsenmedia.com/cellphonesummit/cellphone.html

Cell Phone Sampling

The rise of personal cell phone ownership in many industrialized countries and, more important, the increase in the number of people who can be contacted only via cell phone poses some challenges to traditional telephone surveys. Some of the sampling techniques used for selecting traditional landline (wired) telephone samples still apply when selecting cell phone samples. There are, however, specific characteristics of the cell phone that impact frame construction and sample selection that should be incorporated into designs to maximize yield from cell phone samples. The sampling issues will vary by country as a function of differing cell phone penetration rates, numbering taxonomies, and local market conditions, including technology and plan attributes. Designs for cell phone sampling and weighting, along with a general consensus for their use in practice, are currently continuing to emerge within the survey research community. Based on a query of cell phone systems worldwide, it does appear that the cell phone situation in the United States has a tendency for more complexities. The solutions for other countries may be much simpler versions of these designs.

The New Phone Subscriber Population

The cell phone subscriber population is expanding worldwide and is rapidly changing telephone systems and how people communicate within them. In some countries, the ratio of cell phone subscribers to total residents is quickly reaching a 1:1 ratio. Only 15 years ago, these ratios were in the range of 1:20 to 1:10. Table 1 summarizes the penetration rate of cell phones in selected countries (unadjusted for multiple cell phone ownership) collected by the International Telecommunication Union in 2005. Comparisons between countries should be made carefully due to variations in age distributions within different countries, since age is associated with cell phone ownership. The table gives an idea

Table 1 Cell phone penetration rates by selected countries, 2006

Australia	97
Austria	103
Belgium	93
Canada	53
Denmark	107
Finland	108
France	85
Germany	102
Greece	100
Hong Kong	131
Italy	124
Japan	79
Netherlands	97
Portugal	116
Russia	84
Spain	106
Sweden	105
Turkey	71
Taiwan	97
U.K.	116
U.S.	77

Source: International Telecommunication Union (2006).

of potential undercoverage biases that may result in samples of landline phones that exclude cell phones.

The percentage of cell phone numbers (CPNs) to total inhabitants generally overestimates the number of unique users as reflected by the reality that multiple numbers may be used by a single subscriber. Thus a sampling frame of CPNs may have a problem of multiple listings for some individuals, thereby increasing the probability of selection for those subscribers with multiple CPNs. Another phenomenon that has direct impact on telephone surveys in general is masked in Table 1: In many countries the number of people dismissing a landline or not having one in the first place is also rising. Currently, it is not unrealistic to predict that, in the near future, in some countries everyone could potentially be reached more easily via a cell phone than by a landline phone.

Diversification of Telephone Sampling Frames

As a result of the new presence of cell phone subscribers, the telephone subscriber universe as we know it is changing and can be best described in four parts: (1) cell phone only (CPO), (2) landline only (LLO), (3) cell and landline (C&L), and (4) no phone service of any kind (NPS), as depicted in Figure 1.

In Table 2, the distribution of the population within each of these four subsets is provided for several countries. These data were obtained via nationwide probability samples using face-to-face interviews. A common theme among industrialized countries is the continued rise in the number of inhabitants who fall into the "cell phone only" category; this increase poses threats to undercoverage bias for traditional telephone surveys that typically sample households via random-digit dial samples from frames consisting of only landline numbers (LLNs).

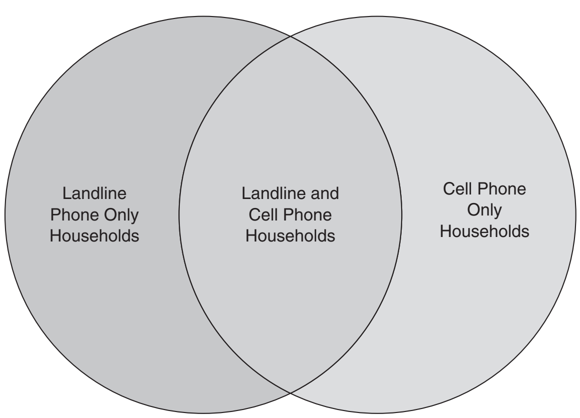

Figure 1 New telephone landscape

Table 2 Household landline and cell phone ownership in selected countries

Country	*Cell Only*	*Cell and Landline*	*Landline Only*	*No Phone*	*Month/Year*
Canada	5.1	61.3	32.4	1.2	12/2006
Finland	52.2	44.3	3.1	0.4	08/2006
France	16.7	61.6	20.8	1.0	07/2006
U.K.	9.0	84.0	7.0	1.0	07/2007
U.S.	14.0	72.3	12.3	1.4	06/2007

In response to the diversification of the telephone universe, the researcher wishing to conduct telephone sampling is now faced with two key questions:

1. Is the amount of undercoverage in a probability sample selected from a frame of only LLNs acceptable? A related question that is usually asked in making the decision regarding the impact of the undercoverage of CPO households is, "How different are CPO households with respect to survey variables?"
2. Is the amount of undercoverage in a probability sample selected from a frame containing only CPNs acceptable? In this case, a related question is, "How different are LLO households for the survey variables of interest?"

In the case where neither single-frame approach (i.e., using a frame of only LLNs or a frame of only CPNs) will produce acceptable estimates (i.e., minimal undercoverage bias, etc.), does the researcher need to employ a dual-frame sampling design consisting of independent samples selected from available landline as well as cell phone number banks (i.e., collections of phone numbers that are grouped according to a combination of area code [United States], prefix, and suffix; a "10,000-bank," for example, represents numbers that have the same area code and prefix [e.g., 999-888-XXXX])?

Cell Phone Sampling Designs

In response to these two scenarios, at least two types of sampling designs can be used to select a cell phone sample, including those involving the selection of only CPNs and those designs that select a cell phone sample in conjunction with a landline sample. For the first case, a sample of cell phones can be selected from a frame constructed using CPNs that have been identified via area code and prefix combination (United States) or simply via prefix (Europe). Selection strategies such as systematic or stratified random sampling (stratified by provider, area code, etc.) can be used with the cell phone number frame.

For the second case, the researcher can employ a dual-frame sample in which a sample of cell phone numbers is selected from the cell phone frame and a second sample of landline numbers is selected from the landline frame. The sampling plans within these two frames can be similar or different. For example, list-assisted sampling plans are generally more efficient for landline phones but may not be a useful design strategy for cell phones, as many countries do not have published lists of working CPNs. More auxiliary information may be available for landline numbers (i.e., corresponding addresses), so stratified random sampling designs may be more feasible for landlines. However, stratifying the cell phone frame by provider or sorting the selected sample by provider may be a very efficient way to incorporate provider variations or add to the efficiency of calling designs once the sample of CPNs is selected.

Regardless of the sampling design used for selecting a cell phone sample, selection of multiple members from a single household is possible for those individuals who live in households with multiple cell phone subscribers. Depending on the survey outcome of interest, the clustering of people by household within the sample may slightly inflate the design effect (*deff*), with the degree of the inflation being a function of the sampling design, the overall penetration rate, and the sample size. In contrast, samples of landline numbers typically use techniques such as the "latest birthday" to randomly select one and only one member from the household for inclusion in the sample. However, a similar clustering effect could happen in landline samples if multiple numbers (and adults) were selected for a single household.

Regardless of the single- or dual-frame sampling designs used to select the sample of CPNs (and LPNs), standard weighting techniques consistent with the chosen design can be used to derive estimates appropriate for inference to each frame. Because the initial sampling units for cell phones are usually people—whereas for landlines it is households—it is important to adjust the weights of these estimators so inference can be made about a common unit. For inference about households, it will be necessary to adjust the initial sampling weights for the number of cell phones or landline phones per household; for person-level inference, additional adjustments incorporating the number of users per cell or landline will be necessary. For dual-frame estimators, these adjustments are typically done separately for each sample drawn from each respective frame.

Traditional dual-frame estimators are derived using separate unbiased estimates for CPO and LLO based on the sample of CPNs and LLNs, respectively, along with a composite estimate that optimally combines the two estimates of the C&L overlap. Treating the dual-frame sample data as though it were from one larger sample, researchers can derive "single-frame estimators" that do not have a separate and explicit component for the overlap.

The single-frame estimator does not make use of frame sizes (which in the case of telephone sampling should be known—that is, banks from which samples are drawn have a fixed size, usually either 10,000 or 1,000), nor does it take advantage of the relative efficiency of the sampling designs used for selecting samples in the two frames. The single-frame estimator can incorporate the known frame sizes via raking ratio estimation or regression. While the form of the estimator does not have a component that comes directly from the "overlap" of people or households from the cell and landline frames, it does require knowledge of the inclusion probabilities in each of the respective frames. For example, for each person or household in the cell phone sample who has at least one landline number, it is necessary to determine the probability for being included in the landline sample, and vice versa. In practice, this amounts to computing the number of both landlines and cell phones that could be used to contact the person or household for all those households or people who fall into the C&L domain. Device grids are a novel tool that can be used in practice as a basis for collecting data from sampled numbers on the number and type of phone devices attached to the household as well as the number of people in the household who use each device. These data then form the basis of person-level weights to be used for person-level inference from single-frame estimators.

The dual-frame estimators avoid the need to compute sample inclusion probabilities for the second "recharged" during the course of ownership). Pre-paid plans sometimes imply multiple cell phone devices per person in the population of interest. For example, in Italy, where a bulk of the plans would be considered pre-paid, the penetration rate for cell phone subscribers was 124% (or 1.24:1) as seen from Table 1. A study conducted in 2002 estimated that upward of one fourth of Italian subscribers owned more than one cell phone number. While the multiplicity of devices per person certainly increases the overall hit rate for samples of cell phone subscribers, it does have implications for the effective sample size of unique subscribers for any given randomly selected sample of CPNs from a CPN frame. As people move from using one cell phone to the other, temporary usage or transitional usage patterns may also impact the number of cell phones with unknown eligibility (i.e., ring–no answer), or a continuous string of only voicemails). In general, pre-paid plans have either no long-term commitments or have generally shorter contract periods than post-paid plans. In the United States, typical post-paid plans have contract periods between 1 and 3 years. These plans tend to make the sampling frame of CPNs more stable over a given study period, but it is possible for CPNs to remain active while the subscribers attached to those numbers change, resulting in potentially ambiguous call outcomes over longer study periods. Experience suggests that shorter field periods for making dialing attempts to reach the user(s) of the CPN, as compared to longer periods for typical landline phone surveys, may be more cost-effective for cell phone sample surveys.

Within contract types there are various plan attributes that may vary within and among providers. For example, in countries such as Canada, the United States, and Hong Kong, cell phone subscribers pay for incoming calls; in many European countries, Japan, and Australia, subscribers receive incoming calls for free. Usually, cell phones worldwide have some type of caller identification that shows the number or programmed name of the caller. This feature, along with the trend of having the called party pay, has a potential impact on the cell phone user's propensity to answer a survey call and also on the general response rate of sample surveys using CPNs.

Cell Phone Sampling in Practice

While limited information is available from just a cell phone number, in the United States the area code or prefix of a cell phone number conveys some level of geographic specificity, and this portion of the phone number can be linked to a larger exchange database to acquire the name of the provider, which can then be used by the researcher as additional stratification variables, namely provider. Also, some providers offer more localized services with free incoming calls or more pre-paid plans that may be associated with a specific demographic target of interest (e.g., younger, college-age subscribers). Of course, stratifying the sample frame by provider allows researchers flexibility in having different sampling plans with the potential to maximize coverage across geographic areas (served sometimes exclusively by some providers, especially in rural areas) and age groups.

At this point in practice there is little evidence to suggest that stratifying cell phone samples by provider increases the accuracy of resulting estimators. In general, however, if questions relating to the usage of technology-related options of cell phone plans, such as Internet, text messaging, or photo exchange, are of interest, then variations in provider offerings may be at a level that provider stratification may improve the overall efficiency of the estimates. Perhaps more useful at this point in the evolution of cell phone practice would be a design that includes a post-stratification of the sample by provider prior to subscriber contact. Much like responsive call designs, provider information can be used to screen numbers for nonworking status using text messaging interfaces available from provider Web sites as well as to design optimal calling schedules based on the off-peak hours generally offered by the providers. In general, calling rule strategies that can take advantage of cell phone provider plan attributes, such as peak and off-peak call time differences or uniform text messaging options or other technologies that are offered to a majority of subscribers from a particular provider, may be more efficient in terms of overall survey yield. As another example, the time intervals associated with peak and off-peak usage vary more across than within provider. For a given plan, subscribers are generally allocated fewer peak time minutes than off-peak time minutes. However, common times for survey researchers to contact sampled cell phone subscribers generally coincide with peak time intervals. In contrast to calls made during peak times, those made during off-peak times do not generally pose a threat of additional or higher costs for the subscriber. Thus "time called" may be a predictor for response in some cases where the called party pays—in these cases, it may be important to vary the day and time called to include peak and off-peak time intervals and weekdays and weekends. On the other hand, some cell phone providers either offer plans for free incoming calls or simply do not charge for incoming calls; such cell phone numbers could be called first in a provider-assisted call design, for example.

Regardless of the design or calling strategy, there are some instances in which disposition codes for cell phones may need to be modified to better describe the different landscape. For example, the proliferation of family plans in the United States is creating multiple cell phones per household. Many of the cell phones within a household will be registered to adults but used primarily or exclusively by children under 18. The disposition "ineligible-underage" is not commonly encountered in landline (household) samples and may need to be added to cell phone sample call disposition codes to more precisely describe the larger "ineligible" category. Rather than imply that there is no adult 18 years or older in the household, this disposition when used with cell phones would imply that the primary user is under 18 years of age and is thus ineligible for surveys of the adult population. While family plans are becoming more popular, there is also some current evidence to support a small degree of sharing of cell phones within households in the United States. In particular, some studies have suggested that cell phone sharing may occur more frequently between adult and child; with many surveys excluding children, the number would either be ineligible or the adult would be selected if an age-appropriate screener were included in the protocol. At this point there is no overwhelming evidence to suggest that within-household selection techniques are required for cell phone samples. However, as the penetration of cell phones increases and as the number of households having multiple cell phones per household increases, these types of selection techniques may become necessary.

The practice of telephone survey research is transitioning in response to the proliferation of cell phone use worldwide. While many of the survey research methods described are currently being used in conjunction with sample surveys of CPNs, it should be noted that general consensus for "best practices" for sampling designs, calling strategies, and weighting algorithms are at best in the experimental phases. As the cell phone landscape continues to evolve within the United States and worldwide, additional information will become available to confirm and possibly reform the current methods.

Trent D. Buskirk and Mario Callegaro

See also Calling Rules; Cell Phone Only Household; Design Effect (*deff*); Dual-Frame Sampling; Federal Trade Commission (FTC) Regulations; Hit Rate; Latest-Birthday Selection; List-Assisted Sampling; Mitofsky-Waksberg Sampling; Number Portability; Prefix; Suffix Banks; Telephone Surveys; Weighting; Within-Unit Selection

Further Readings

Brick, J. M., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, *70*, 780–793.

Callegaro, M., & Poggio, T. (2004). Espansione della telefonia mobile ed errore di copertura nelle inchieste telefoniche [Mobile telephone growth and coverage error in telephone surveys]. *Polis*, *18*, 477–506. English version retrieved March 24, 2008, from http://eprints.biblio.unitn.it/archive/00000680

Callegaro, M., Steeh, C., Buskirk, T. D., Vehovar, V., Kuusela, V., & Piekarski, L. (in press). Fitting disposition codes to mobile phone surveys: Experiences from studies in Finland, Slovenia, and the United States. *Journal of the Royal Statistical Society*, Series A (Statistics in Society).

International Telecommunication Union. (2006). *World telecommunication indicators database* (9th ed.). Geneva: Author.

Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame rdd surveys. *Public Opinion Quarterly*, *71*(5), 750–771.

Kuusela, V., Callegaro, M., & Vehovar, V. (2007). Mobile phones' influence on telephone surveys. In M. Brick, J. Lepkowski, L. Japec, E. de Leeuw, M. Link, P. J. Lavrakas, et al. (Eds.), *Telephone surveys: Innovations and methodologies* (pp. 87–112). Hoboken, NJ: Wiley.

Lavrakas, P. J., & Shuttles, C. D. (2005) Cell phone sampling, RDD surveys, and marketing research implications. *Alert!*, *43*, 4–5.

Lavrakas, P. J., Shuttles, C. D., Steeh, C., & Fienberg, H. (2007). The state of surveying cell phone numbers in the United States: 2007 and beyond. *Public Opinion Quarterly*, *71*(5), 840–854.

Lohr, S., & Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, *95*, 271–280.

Steeh, C., Buskirk, T. D., & Callegaro, M. (2007). Using text messages in U.S. mobile phone surveys. *Field Methods*, *19*, 59–75.

Cell Suppression

Under certain circumstances, it is considered necessary to withhold or suppress data in certain cells in a published statistical table. This is often done when particular estimates are statistically unreliable or when the information contained could result in public disclosure of confidential identifiable information. Suppression for reasons of statistical reliability involves consideration of sampling error as well as the number of cases upon which the cell estimate is based. Suppression to avoid the disclosure of confidential information in tabular presentations involves many additional considerations.

Cell suppression may involve *primary* suppression, in which the contents of a sensitive cell are withheld; or if the value for that cell can be derived from other cells in the same or other tables, *secondary* or *complementary* suppression. In the latter instance, the contents of nonsensitive cells as well those of the sensitive cells are suppressed. Sensitive cells are identified as those containing some minimum number of cases. In an establishment survey, for example, a cell size of 2 would be regarded as sensitive because it could reveal to one sample establishment (included in the tabulation and knowing its contribution to an estimate reported in the table) the value of a variable reported by another establishment known to have participated in the survey. Often, the minimum cell size for suppression is considerably higher than 2, depending upon such factors as total sample size, sampling ratio, and potential harm to survey participants resulting from disclosure.

Once sensitive cells have been identified, there are some options to protect them from disclosure: (a) restructure the table by collapsing rows or columns until no sensitive cells remain, (b) use cell suppression, (c) apply some other disclosure limitation method, or (d) suppress the entire planned table.

When primary and complementary suppressions are used in any table, the pattern of suppression should be audited to check whether the algorithms that select the suppression pattern permit estimation of the suppressed cell values within "too close" of a range. The cell suppression pattern should also minimize the amount of data lost as measured by an appropriate criterion, such as minimum number of suppressed cells or minimum total value suppressed. If the information loss from cell suppression is too high, it undermines the utility of the data and the ability to make correct inferences from the data. Cell suppression does create missing data in tables in a nonrandom fashion, and this harms the utility of the data.

In general, for small tables, it is possible to select manually cells for complementary suppression and to apply audit procedures to guarantee that the selected cells adequately protect the sensitive cells. However, for large-scale survey publications having many interrelated, higher-dimensional tables, the selection of a set of complementary suppression cells that are optimal is an extremely complex problem. *Optimality* in cell suppression is achieved by selecting the smallest number of cells to suppress (to decrease information loss) while ensuring that confidential information is protected from disclosure.

Stephen J. Blumberg

See also Confidentiality; Disclosure Limitation

Further Readings

Gonzalez, J. F., & Cox, L. H. (2005). Software for tabular data protection. *Statistics in Medicine*, *24*, 659–669.

Klein, R. J., Proctor, S. E., Boudreault, M. A., & Turczyn, K. M. (2002). Healthy People 2010 criteria for data suppression. National Center for Health Statistics. *Statistical Notes*, no. 24. Retrieved March 24, 2008, from http://www.cdc.gov/nchs/data/statnt/statnt24.pdf

Census

A census is an attempt to list all elements in a group and to measure one or more characteristics of those elements. The group is often an actual national population, but it can also be all houses, businesses, farms, books in a library, cars from an assembly line, and so on. A census can provide detailed information on all or most elements in the population, thereby enabling totals for rare population groups or small geographic areas. A census and a sample survey have many features in common, such as the use of a questionnaire to collect information, the need to process and edit the data, and the susceptibility to various sources of error. Unlike a sample survey, in which only a subset of the elements is selected for inclusion and enumeration, a census generally does not suffer from sampling error. However, other types of errors may remain. The decision to take a census versus a sample survey—if not mandated by statute—is often based on an assessment of the coverage, cost, errors in the data, and other qualitative factors.

Aspects of a census include the types and historical purposes for taking a census, its statistical properties, the differences between a census and a sample survey, and errors that can occur in a census.

General Background

Perhaps the most well-known type of census is one that enumerates the population or housing characteristics of a specified country or other politically defined region. Others measure the output in a specified sector of the economy, such as agriculture, transportation, manufacturing, or retail sales. These censuses are typically authorized and funded by the central government of the region covered.

Censuses were first conducted hundreds (Canada, Sweden) and even thousands (China) of years ago in some parts of the world. In many countries, a census is repeated in a fixed cycle, often every 5th (the United Kingdom, Canada, Australia, New Zealand) or 10th (Portugal, Spain, Italy, Poland, Turkey) year. In the United States, the census of population and housing has been conducted every 10th year, beginning in 1790. The U.S. economic census is taken every 5th year.

Historically, the purpose of the census has varied. At first, governing bodies wanted to know the number of people for assessing taxes or determining the number of men eligible for the military. Currently, governments use census data to apportion their legislative bodies, set boundaries for political districts, distribute government funds for social programs, track the nation's economy, measure crops to predict food supplies, and monitor people's commute to work to determine where to improve the region's infrastructure. As a by-product, census lists of households, businesses, or farms are often used as frames for surveys or follow-up studies. Further, the detailed information collected in the census allows for more efficient sample designs and improved estimation in the surveys.

Content and Mode of Collection

The content of a census form can range from a few basic questions to many detailed questions. Indeed, the same census may combine the two approaches. In recent decades, in the U.S. Census of population and housing most households received a "short form" limited to the names, ages, and a few other characteristics of the people living in the household. At the same time, a sample of about 1 in 6 U.S. households received a "long form" that solicited the basic information as well as more detailed data on the residents' demographic and educational background, the housing unit's physical size and structure, and other characteristics. Plans for the U.S. Census in 2010 call for only a short form. The detailed data formerly solicited in the long-form census are now collected in the American Community Survey, a large survey conducted by the U.S. Census Bureau designed to produce estimates at the county level every year. In an economic census, dozens of different forms may be used to tailor the questions to specific types of business.

Traditionally, census takers went door to door asking questions, an approach still used in many countries, especially in the developing world. In the developed world, one or several modes of enumeration may be used. People or businesses are often contacted by mail or in person, perhaps by telephone if a current number is available. When no response is received from a mailing, a census representative may be sent to a housing unit or establishment to follow up. Where feasible, especially when canvassing businesses, an electronic questionnaire might be provided on a disk. In some censuses, respondents may be encouraged to reply via the Internet.

Alternative or combination approaches can be used to solicit or collect data. As an example, in the U.S. Census of Retail Trade in 2002, all of the larger establishments and a sample of the smaller ones were mailed a complete questionnaire. For the smaller firms not selected into the sample, the basic economic information was collected through available tax records. Such an approach can lessen the reporting burden of the respondents and, in some cases, provide valuable auxiliary data. However, combining alternative methods of data collection usually requires an examination of several key aspects: the coverage of the population, differences in the definitions of data items, the consistency of information collected via different modes, and the accuracy of the data.

To Take a Census or a Sample Survey?

A census generally attempts to collect information on all eligible elements in a defined population, while a sample survey pre-selects a subset of elements for inclusion. But it is doubtful whether any census has ever successfully captured all elements, for reasons involving frame deficiencies, census procedures, the cooperation of respondents, or other issues. While a census may produce almost complete coverage, there are also advantages to taking a sample survey. To start, taking a census requires extensive planning and complex operations. In making contact with only a fraction of the population, a sample survey usually imposes a burden on many fewer respondents and costs much less to complete.

Some costs—questionnaire materials, mailing charges, interviewer salaries—tend to be proportional to the size of the canvassed population. Other costs can escalate with the size. For example, when planning for a large-scale census, one might have to hire and train two or three times as many interviewers as will be needed during the census, because many will drop out or be discharged before the census is completed. With a sample survey, because of the smaller scale of the operation, one can better control the hiring and training of interviewers and thus lower costs. For repeated surveys or when several surveys are run out of the same field office, interviewers who work on one survey may be used on other surveys when their schedules permit, taking advantage of experience and reducing training costs.

The decision to take a census or a sample survey is at times a trade-off between the breadth of detail and the currency of the information. Often, only a census can produce useful information for rare populations or small geographic areas. For example, the U.S. Census produces data for the population classified by age, race, and Hispanic identity for each block in the country. No survey could possibly produce such information. Yet, in a census, data are generally collected at one point in time and can take months or years to process and disseminate. When it is released, that information may have to suffice until the next census is completed and processed. On the other hand, a survey can be taken at much more frequent intervals—perhaps on a monthly, quarterly, or annual basis—but might collect only a subset of the information captured in the census.

Errors in a Census

While the results from a census typically do not suffer from sampling error—those errors introduced by canvassing only a sample of the entire population—censuses are susceptible to the nonsampling errors found in sample surveys. A common problem is missing data, such as unit nonresponse (when no usable data are obtained for a population element) or item nonresponse (when only a portion of a response is usable), due to failure to reach the respondent or the respondent's unwillingness or inability to provide information.

Nonsampling errors can arise in various ways. Respondents can misinterpret questions on the census form, especially if the questions are vague or too complex. Errors may be introduced when respondents must estimate the quantity requested on the questionnaire. When conducting a personal interview, the behavior of a census field representative can influence the responses. Other sources of nonsampling errors include coverage problems (undercoverage or overcoverage of the target universe), processing errors, and mistakes recording or keying data. For example, census data describing industry or place of work must be coded to be useful. But coding can introduce both random and systematic errors into the census results.

To address nonsampling errors, statistical procedures are sometimes applied. For example, to treat unit or item nonresponse, a missing item might be replaced by the item's value from a respondent whose characteristics are similar to those of the nonrespondent. Inserting values for missing items on a questionnaire is called "imputation."

In a sample survey, sampling error generally decreases as the size of the sample increases. But any systematic biases introduced in a census process or operation generally are not eliminated—even though the entire population is canvassed or targeted. Estimating the size of nonsampling errors requires follow-up studies or data from independent sources. As a result, the level of nonsampling error in a census is generally not known or published.

Because conducting a sample survey is a much smaller operation than taking a complete census, nonsampling errors can sometimes be contained better in surveys. A greater proportion of the allotted time and budget can be spent obtaining responses, eliminating sources of error, and improving the quality of the data. Consequently, at times survey results can be more accurate than census results. Still, by attempting to cover the entire population, a census retains advantages over a sample survey. As mentioned previously, a census provides direct summary statistics for the characteristics of small areas or domains. With a sample survey, indirect methods or models are often required to produce small-area estimates when the size of the sample falling in the area or domain is too small. Such procedures are susceptible to errors when the models are specified incorrectly.

Statistical procedures—including probability sampling—are often used while a census is being taken and after its completion. For example, quality control measures can be applied in a sample of regions to monitor operations and determine whether procedures are being followed as specified. After the enumeration, to measure the coverage or accuracy of the census, a sample of areas or domains may be selected and examined in greater detail. Data obtained from re-interviews or administrative records can be used to produce estimates of the total number of census omissions or erroneous enumerations in the entire population or in subgroups.

Patrick J. Cantwell

See also American Community Survey (ACS); Confidentiality; Coverage Error; Imputation; Interviewer Effects; Missing Data; Mode of Data Collection; Nonresponse; Nonsampling Error; Sampling Error

Further Readings

Anderson, M. J. (1988). *The American census: A social history.* New Haven, CT: Yale University Press.

Anderson, M. J., & Fienberg, S. E. (1999). *Who counts? The politics of census-taking in contemporary America.* New York: Russell Sage Foundation.

Hansen, M. H., Hurwitz, W. N., & Bershad, M. A. (1961). Measurement errors in censuses and surveys. *The Bulletin of the International Statistical Institute*, *38*, 359–374.

Kish, L. (1979). Samples and censuses. *International Statistical Review*, *47*, 99–109.

Kish, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, *14*, 1, 31–46.

United Nations Statistics Division, World Population and Housing Census Programme: http://unstats.un.org/unsd/demographic/sources/census/default.aspx

Certificate of Confidentiality

In order to collect sensitive information, researchers need to be able to ensure for themselves that identifiable research data will remain confidential and assure respondents that this is the case. However, neither legislatures nor courts have granted researchers an absolute privilege to protect the confidentiality of their research data. Despite this, there are several federal statutory mechanisms that can be helpful. In some cases researchers can obtain legal protection for the confidentiality of research data through a federally issued Certificate of Confidentiality as authorized by the Public Health Service Act § 301 (d), 42 U.S.C § 241(d):

> The Secretary may authorize persons engaged in biomedical, behavioral, clinical, or other research (including research on mental health, including research on the use and effect of alcohol and other psychoactive drugs) to protect the privacy of individuals who are the subject of such research by withholding from all persons not connected with the conduct of such research the names or other identifying characteristics of such individuals. Persons so authorized to protect the privacy of such individuals may not be compelled in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify such individuals.

Certificates of Confidentiality allow the investigator and others who have access to research records to refuse to disclose identifying information on research participants in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level. Certificates of Confidentiality may be granted for studies collecting information that, if disclosed, could have adverse consequences for subjects or damage their financial standing, employability, insurability, or reputation (such as drug use, sexual behavior, HIV status, mental illness).

Research need not be federally supported to be eligible for this privacy protection. Certificates of Confidentiality are issued by various Public Health Service component agencies, the Food and Drug Administration, the Health Resources and Services Administration, and the National Institutes of Health. Researchers are expected to inform subjects in the consent form about the Certificate of Confidentiality protections and the circumstances in which disclosures would be made to protect the subject and others from harm (such as suicidal intention, child abuse, elder abuse, intention to harm others) and certain types of federal audits.

There is very little legal precedent considering the scope of the protections afforded by Certificates of Confidentiality. However, in at least one case from 1973 (*People v. Newman*), a New York state court of appeals found that a certificate provided a substance abuse program with a proper basis for refusing to turn over the names of program participants.

There are other types of legal protection available for some federally funded research. The privacy of research subjects in Department of Justice–funded research is protected by statute—42 U.S.C. Section 3789g. Similarly, the privacy of research subjects in Agency for Health Care Quality and Research–funded research is protected by a statute 42 U.S.C. Section 299a-1(c) titled "limitation on use of certain information." For these studies, Confidentiality Certificates are not appropriate. All researchers collecting sensitive data as part of projects under the jurisdiction of an institutional review board will need to work closely with their board and also may require legal counsel.

Sandra H. Berry

See also Ethical Principles; Institutional Review Board; Survey Ethics

Further Readings

Merewitz, S. G. (2001). *Agency for Healthcare Research and Quality, Statutory confidentiality protection of research data collected with AHRQ support.* Retrieved January 4, 2007, from http://www.ahrq.gov/fund/datamemo.htm

National Institute of Justice. (2007). *Human subjects protection.* Retrieved March 26, 2008, from http://www.ojp.usdoj.gov/nij/funding/humansubjects

National Institutes of Health, Office of Extramural Research. (n.d.). *Certificates of Confidentiality kiosk.* Retrieved January 4, 2007, from http://grants1.nih.gov/grants/policy/coc/index.htm

Traynor, M. (1996). Countering the excessive subpoena for scholarly research. *Law and Contemporary Problems*, *59*, 119–148. Retrieved June 2, 2008, from http://www.law.duke.edu/shell/cite.pl?59+Law+&+Contemp.+Probs.+119+(Summer+1996)

CHECK ALL THAT APPLY

The check-all-that-apply question format presents respondents with multiple response options to a single question, as shown in Figure 1.

What race or races are you? (Please check all that apply)

___ Asian
___ Black
___ Native American
___ Pacific Islander
___ White
___ Other (Please specify:_________________)

Figure 1 Check all that apply

In response to the question, the respondents are instructed to select as many of the response options as are perceived to apply to them. Although the check-all-that-apply question format is commonly used in survey questionnaires, research has shown that it can result in a less than optimal response strategy by respondents and may be especially sensitive to primacy effects when the question is asking about past experiences, behaviors, or attitudes.

When evaluating a list of response options to a check-all-that-apply question, respondents may strive for satisficing and burden avoidance. For example, respondents may select only the first of several reasonably acceptable response options and fail to adequately consider the remaining response options before proceeding to the next question. Because of this, some researchers believe it is important to deploy several versions of a check-all-that-apply question, with the response options listed in different orders that are randomly assigned to different respondents, so as to scramble the order of the list of response options across the entire sample.

The check-all-that-apply question format is distinct from the forced choice format (e.g., a list of Yes/No response options). In the forced choice format, respondents are asked to evaluate each forced choice response option individually before moving on to the next. The literature suggests that this difference may result in respondents following divergent cognitive approaches in responding to the forced choice format versus the check-all-that-apply format. In particular, respondents may show more careful consideration and greater cognitive processing of each response option in the forced choice format, while selecting only the first few of several response options that apply in the check-all-that-apply format. Research has shown that in addition to a primacy effect associated with the check-all-that-apply format, the difference between the two formats may result in a higher average number of response options selected per respondent in a forced choice question than in a comparable check-all-that-apply question.

While the addition of the "No" category in the forced choice format should provide greater discrimination when compared to the check-all-that-apply format (which lacks an explicit "No" category), research also has shown that, without adequate instruction, respondents may treat a forced choice format in self-administered questionnaires as Check All That

Apply. This occurs when respondents correctly select the "Yes" category for positive responses but fail to select the "No" category for negative responses. As a result, the data can be difficult to interpret. Blank responses may either be intended as a negative response, a not applicable response, or simply an undecided, don't know, or a missing response.

The check-all-that-apply question format is commonly used in self-administered paper-based and Internet surveys. It is less well suited to telephone surveys and consequently is rarely used in that mode. In interviewer-administered in-person surveys, use of the check-all-that-apply format should be paired with the use of a show card displaying the choices to the respondent. In multi-mode surveys, there has been a tendency to pair a check-all-that-apply question in a self-administered questionnaire with a forced choice version in a telephone interview. However, considering the findings in the literature that show that respondents do not treat the two question formats similarly, converting a check-all-that-apply question from a self-administered questionnaire to a forced choice format for use in a telephone interview may not be an optimal approach.

Adam Safir

See also Forced Choice; Primacy Effect; Questionnaire Design; Response Order Effects; Satisficing; Show Card

Further Readings

Krosnick, J., & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219.

Rasinski, K., Mingay, D., & Bradburn, N. (1994). Do respondents really "mark all that apply" on self-administered questions? *Public Opinion Quarterly*, *58*(3), 400–408.

Smyth, J., Dillman, D., Christian, L., & Stern, M. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly*, *70*(1), 66–77.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire construction.* San Francisco: Jossey-Bass.

CHI-SQUARE

The chi-square (χ^2) is a test of significance for categorical variables. Significance tests let the researcher know what the probability is that a given sample estimate actually mirrors the entire population. The chi-square can be used as a goodness-of-fit test, in univariate analysis, or as a test of independence, in bivariate analysis. The latter is the most generally used. In this case, the test measures the significance of the relationship between two categorical variables, representing the first step toward bivariate analysis. For example, if a survey researcher wanted to learn whether gender is associated with an attitude (negative or positive) toward the U.S. involvement in Iraq, chi-square is the simplest significance test to consider to investigate whether or not there are reliable gender-related differences in these attitudes (see Table 1).

The logic behind the chi-square is to calculate the distance between the observed frequencies within the contingency table and the condition of statistical independence (i.e., the hypothesis of no association or "null hypothesis"). The frequencies that Table 1 would contain in case of no association (the so-called expected frequencies) are calculated by dividing the product of the marginal frequencies (row and column) of each cell by the sample size. The greater the distance between the observed frequencies and the expected frequencies, the higher is the chi-square. This is the formula:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e},$$

where f_o represents the observed frequencies and f_e are the expected frequencies. If the value of the chi-square is 0, there is no association between the variables. Unfortunately, the chi-square has no maximum, and this makes its interpretation not intuitive.

In order to interpret the value obtained, the researcher must first calculate the degrees of freedom (*df*) of the contingency table, multiplying the number of the rows minus 1 by the number of the columns minus 1. Second, given the values of chi-square and *df*, he or she has to search for the corresponding value of *p*-level. This value can be located on the chi-square

Table 1 Example of contingency table for chi-square analysis (frequency counts)

Support/Oppose U.S. Involvement in Iraq	*Female*	*Male*	*Total*
Support	**170**	**200**	370
Oppose	**250**	**150**	400
Total	420	350	770

distribution table, usually reported in most handbooks of statistics, or calculated through statistical software such as Statistical Package for the Social Sciences (SPSS) or SAS.

The *p*-level is the crucial figure to consider when evaluating the test. This is the actual value that indicates the significance of the association. It says, in short, how probable it is that the relationship observed in the survey data is due to mere sampling error. The chi-square test must be used cautiously. First, the researcher should have a probability sample whose size is ≥ 100. Second, since the chi-square statistic is sensitive to the sample size, the researcher cannot compare the chi-square values coming from different samples. Third, researchers should be careful that the expected values in the contingency table are not too small (≤ 5), because the chi-square value will be heavily biased. Finally, sometimes it makes no sense to calculate the chi-square: for example, when the number of categories of both variables is too high.

In all these cases, the chi-square test should not be separated from the detailed inspection of the contingency table and/or the use of more sophisticated measures. Since the chi-square value is not easily interpretable, other measures have been derived from it, like phi-square, Pearson's C, and Cramér's V. They are not influenced by the sample size and, above all, tend to range from 0 to 1 (this maximum, however, is actually achievable only by Cramér's V), measuring the strength of the association, even when this latter is nonlinear.

Alberto Trobia

See also Contingency Table; *p*-Value; Research Hypothesis; SAS; Statistical Package for the Social Sciences (SPSS)

Further Readings

Blalock, H. M. (1979). *Social statistics*. New York: McGraw-Hill.

Bohrnstedt, G. W., & Knoke, D. (1994). *Statistics for social data analysis*. Ithaca, NY: Peacock.

CLOSED-ENDED QUESTION

A closed-ended survey question is one that provides respondents with a fixed number of responses from which to choose an answer. It is made up of a question stem and a set of answer choices (the response alternatives). When administered by a survey interviewer, a closed-ended question is expected to be read exactly as written to the respondent, along with the full set of response alternatives. The set of answer choices must fulfill two properties: they must be (1) mutually exclusive and (2) exhaustive. In being mutually exclusive, no two answers can overlap in conceptual meaning. In being exhaustive, the answer choices must cover all logically possible answers for the question.

The following example of a closed-ended question has answers that are neither mutually exclusive nor are they exhaustive:

> *How many times in the past 30 days have you entered a grocery store?*
> *(a) 1–5 (b) 6–10 (c) 11–15 (d) 15 or more*

In the example, a respondent who entered a grocery store 15 times in the past 30 days would not know if she or he should choose response (c) or (d), because the two are not mutually exclusive, as both contain the number 15. A respondent who never entered a grocery store in the past 30 days should answer "0," but the response choices do not include that answer and thus they are not exhaustive of all logically possible answers.

With interviewer-administered questionnaires, such as those used in face-to-face and telephone surveys, closed-ended questions typically are constructed so that the interviewer can code a "Don't know/Uncertain" (DK) response when that is appropriate for a given respondent. They also typically include a "Refused" (RF) response choice for the interviewers to code when a given respondent refuses to provide an answer to that question. DK and RF response choices are not provided to the respondent by the interviewer. In self-administered questionnaires, closed-ended questions do not often contain these additional response choices, as their inclusion likely would "open the door" for respondents to avoid providing substantive answers to questions.

Paul J. Lavrakas

See also Balanced Question; Don't Knows (DKs); Exhaustive; Forced Choice; Mutually Exclusive; Open-Ended Question; Precoded Question; Response Alternatives

Further Readings

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Clustering

In broad terms, *clustering*, or *cluster analysis*, refers to the process of organizing objects into groups whose members are similar with respect to a similarity or distance criterion. As such, a cluster is a collection of similar objects that are distant from the objects of other clusters. Unlike most classification techniques that aim to assign new observations to one of the many existing groups, clustering is an exploratory procedure that attempts to group objects based on their similarities or distances without relying on any assumptions regarding the number of groups.

Applications of clustering are many; consequently, different techniques have been developed to address the varying analytical objectives. There are applications (such as market research) in which clustering can be used to group objects (customers) based on their behaviors (purchasing patterns). In other applications (such as biology), clustering can be used to classify objects (plants) based on their characteristics (features).

Depending on the application and the nature of data at hand, three general types of data are typically used in clustering. First, data can be displayed in the form of an $O \times C$ matrix, where C characteristics are observed on O objects. Second, data can be in the form of an $N \times N$ similarity or distance matrix, where each entry represents a measure of similarity or distance between the two corresponding objects. Third, data might represent presumed group membership of objects where different observers may place an object in the same or different groups. Regardless of data type, the aim of clustering is to partition the objects into G groups where the structure and number of the resulting natural clusters will be determined empirically. Oftentimes, the input data are converted into a similarity matrix before objects are portioned into groups according to one of the many clustering algorithms.

It is usually impossible to construct and evaluate all clustering possibilities of a given set of objects, since there are many different ways of measuring similarity or dissimilarly among a set of objects. Moreover, similarity and dissimilarly measures can be univariate or multivariate in nature, depending on whether one or more characteristics of the objects in question are included in calculations. As such, it is impractical to talk about an optimal clustering technique; however, there are two classes of techniques (hierarchical and nonhierarchical) that are often used in practice for clustering.

Hierarchical techniques proceed in a sequential fashion, producing an increasing or decreasing number of nested arrangements of objects. Such techniques can be agglomerative, whereby individual objects start as single clusters and thereafter similar clusters are merged to form progressively fewer larger clusters. As the number of clusters decreases, so do their similarities, eventually leading to the single most dissimilar cluster that includes all objects. In contrast, hierarchical techniques can be divisive, whereby a single cluster of all objects is first partitioned into two clusters of similar objects and thereafter the resulting clusters are further portioned into two new similar clusters. As the number of clusters increases, so do their similarities, eventually leading to the set of most similar clusters that consists of one object per cluster. With hierarchical techniques, the criterion for merging or partitioning interim clusters can be based on the distance (linkage) between their nearest objects, furthest objects, average distance among all objects, or more sophisticated distance measures such as those based on Ward's or Centroid methods. The results of both agglomerative and divisive clustering techniques are often displayed via a two-dimensional graph (tree) called a "dendogram."

Nonhierarchical techniques aim to partition objects into a number of clusters by starting with an a priori set of clusters. Alternatively, such techniques can start the partitioning process based on a set of initial seed points that serve as the nuclei of the emerging clusters. Under either approach, the starting points (initial clusters or seed values) can be chosen in a random fashion to reduce systematic bias. It should be noted that the number of possible clusters of size K that can be formed from O objects can be fairly large (of order $KO/K!$) to allow an exhaustive search for the initial selection.

While there are several nonhierarchical methods of clustering, the method of K-means is the most commonly used technique in practice. This partitioning technique relies on the Euclidean distance between group centroid to measure proximity. Upon formation of the initial K clusters, using either a set of a priori clusters or seed points, the algorithm proceeds by successively assigning each object to the cluster with the nearest centroid. After each reassignment, the centroid points for the donating and receiving clusters are recalculated to identify the structure of the resulting clusters.

Aside from the algorithm chosen for clustering, several guidelines have been developed over the years regarding the number of clusters. While a few of these guidelines rely on visual clues such as those based on sizable change in dendograms, others incorporate formal statistical tests to justify further bisecting of clusters. It has been suggested that visual guidelines can be somewhat ad hoc and result in questionable conclusions. Test-based approaches, on the other hand, might require more distributional conformity than the data can afford.

Mansour Fahimi

See also SAS; Statistical Package for the Social Sciences (SPSS)

Further Readings

Jobson, J. D. (1992). *Applied multivariate data analysis* (Vol. II). New York: Springer-Verlag.

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice Hall.

Seber, G. A. F. (1984). *Multivariate observations.* New York: Wiley.

Cluster Sample

Unlike stratified sampling, where the available information about all units in the target population allows researchers to partition sampling units into groups (strata) that are relevant to a given study, there are situations in which the population (in particular, the sampling frame) can only identify pre-determined groups or clusters of sampling units. Conducive to such situations, a *cluster sample* can be defined as a simple random sample in which the primary sampling units consist of clusters. As such, effective clusters are those that are heterogeneous within and homogenous across, which is a situation that reverses when developing effective strata.

In area probability sampling, particularly when face-to-face data collection is considered, cluster samples are often used to reduce the amount of geographic dispersion of the sample units that can otherwise result from applications of unrestricted sampling methods, such as simple or systematic random sampling. This is how cluster samples provide more information per unit cost as compared to other sample types. Consequently, cluster sampling is typically a method of choice used when it is impractical to obtain a complete list of all sampling units across the population of interest, or when for cost reasons the selected units are to be confined to a limited sample of clusters. That is, feasibility and economy are the two main reasons why cluster samples are used in complex surveys of individuals, institutions, or items.

Operationally, clusters can be defined as collection of units that are geographic, temporal, or spatial in nature. For instance, counties or census blocks often serve as geographic clusters for households sampling; calendar years or months are used for temporal clustering; while boxes of components or plots of land are examples of spatial clusters of objects. Depending on the nature of a study and the extent of heterogeneity among units within each cluster, different numbers of clusters might be needed to secure reliable estimates from a cluster sample. When units within all clusters display the same variability with respect to the measure of interest as the target population as a whole, reasonable estimates can be generated from a small number of clusters. In contrast, when variability is small within but large across clusters, a larger number of clusters of smaller size might be needed to ensure stability.

In spite of feasibility and economical advantages of cluster samples, for a given sample size cluster sampling generally provides estimates that are less precise compared to what can be obtained via simple or stratified random samples. The main reason for this loss in precision is the inherent homogeneity of sampling units within selected clusters, since units in a given cluster are often physically close and tend to have similar characteristics. That is, selection of more than one unit within the same cluster can produce redundant information—an inefficiency leading to higher standard errors for survey estimates.

Kish provided a model for estimating the inflation in standard errors due to clustering. Accordingly, this multiplicative clustering design effect, *deff*, can be estimated by

$$deff = 1 + \rho(\overline{m} - 1).$$

In the preceding formulation, $\overline{m}$ represents the average cluster size and ρ (rho) denotes the so-called intraclass correlation, which is an estimate of relative homogeneity within clusters measured with respect to key analytical objectives of the survey. Obviously, the above effect approaches unity (or no effect) when the average cluster size approaches 1—that is, when the design approaches simple random sampling with no clustering. When ρ becomes exceedingly large due to high correlation between sampling units within clusters, it becomes exceedingly less efficient to select more than one unit from each cluster. Stated differently, even a relatively moderate measure of intraclass correlation can have a sizable inflationary effect on the standard errors when the average cluster size is large.

It should be noted that single-stage cluster sampling is rarely used for selection of the final sampling units. Instead, this methodology is often combined with other sampling techniques to improve the efficiency of the resulting sample. In multi-stage designs, commonly, the first stage consists of stratification of units into similar subsets or those for which reporting is required. It is at the second stage that usually cluster samples are selected within each stratum. Given that sampling with probability proportional to size (PPS) often reduces the standard errors of estimates, cluster sampling provides an ideal framework for this type of sample selection since the number of units in a cluster forms a natural measure of size for the given cluster. In particular, sampling with probabilities proportional to the size of clusters pays big dividends with respect to reducing the error of estimation when the cluster total is highly correlated with the number of units in the cluster.

Mansour Fahimi

See also Area Probability Sample; Clustering; Design Effect (*deff*); Face-to-Face Interviewing; Multi-Stage Sample; Primary Sampling Unit (PSU); Probability Proportional to Size (PPS) Sampling; ρ (Rho); Sampling Frame; Simple Random Sample; Strata; Stratified Sampling; Systematic Sampling; Target Population

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.
Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations.* New York: Wiley.
Scheaffer, R. L., Mendenhall, W., & Ott, L. (2005). *Elementary survey sampling.* Boston: Duxbury.

Cochran, W. G. (1909–1980)

William Gemmell Cochran was an early specialist in the fields of applied statistics, sample surveys, experimental design, observational studies, and analytic techniques. He was born in Rutherglen, Scotland, to Thomas and Jeannie Cochran on July 15, 1909, and he died on Cape Cod, Massachusetts, on March 29, 1980, at the age of 70. In 1927, Cochran participated in the Glasgow University Bursary competition and took first place, winning enough funds to finance his education. After taking a variety of classes, he was awarded an M.A. in mathematics and physics at the University of Glasgow in 1931. He then received a scholarship for a Cambridge University doctoral program, where he studied mathematics, applied mathematics, and statistics. He began his professional career at the Rothamsted Experimental Station in England after being persuaded by Frank Yates to leave Cambridge prior to the completion of his doctorate. Cochran remained at Rothamsted until 1939, working on experimental designs and sample survey techniques, including a census of woodlands with colleague and mentor Yates. During his years at Rothamsted, Cochran remained in touch with R. A. Fisher and was heavily influenced by Fisherian statistics. In his 5 years at Rothamsted (1934–1939), he published 23 papers. Also during his time at Rothamsted, Cochran met and married Betty I. M. Mitchell.

In 1939 Cochran accepted a post in statistics at Iowa State University, where he taught from 1939 to 1946. His task at Iowa was to develop their graduate program in statistics. During his years at Iowa he both served on and chaired the advisory panel to the U.S. Census and published a number of papers on experimental design. Cochran joined Samuel Wilks and the Statistical Research Group at Princeton University in 1943, examining probabilities of hits in naval warfare and the efficacy of bombing raid strategies. Shortly after World War II, he joined Gertrude Cox at the North Carolina Institute of Statistics, where he assisted in developing graduate programs in statistics. Cochran chaired the Department of Biostatistics at Johns Hopkins University from 1949 until 1957. During this time he authored two books, *Sampling Techniques* and (in collaboration with Gertrude Cox) *Experimental Designs*. In 1957 Harvard University established a Department of Statistics and appointed Cochran to head the department. Cochran remained at Harvard until his retirement in 1976.

During his career, Cochran was lauded with many honors. He was the president of the Institute of Mathematical Statistics in 1946, the 48th president of the American Statistical Association in 1953–1954, president of International Biometric Society 1954–1955, and the president of the International Statistical Institute from 1976 to 1981. Cochran was elected honorary fellow of the Royal Statistical Society in 1959, held a Guggenheim Fellowship in 1964, and won the S. S. Wilks medal of the American Statistical Association in 1967. He received honorary doctorate degrees from Johns Hopkins University and the University of Glasgow. From 1974 until his death in 1980, he worked with the National Academy of Sciences' National Research Council panel on incomplete data in sample surveys.

Cochran developed methods for including or excluding an independent variable in multiple linear regression. He also developed the Cochran *Q*-test, used to evaluate two variables measured on a nominal scale. Cochran was the statistical representative for the U.S. Public Health Service research on the effects of smoking on lung cancer. His work as part of the advisory committee provided the surgeon general with proof that lung cancer was directly related to smoking. He also worked on the Kinsey Report on human sexual behavior, on polio research, and on the effects of radiation on Hiroshima victims. He is well remembered for his many agricultural studies such as the yield of cereals, field counts of diseased plants, and the influence of rainfall.

Cochran developed his knowledge of statistics by both studying and working at some of the most prestigious universities. During his lifetime he was involved in diverse research projects and made many important contributions to the field of statistics, not the least of which was establishing statistics departments at several universities. As a teacher, he is remembered for his high expectations for his students, his individuality, and his clarity.

Kathryn A. Cochran and Jody M. Smarr

Further Readings

Anderson, R. L. (1980). A personal tribute to William Gemmell Cochran. *Biometrics*, *36*, 574–578.

Watson, G. S. (1982, March). William Gemmell Cochran 1909–1980. *The Annals of Statistics*, *10*(1), 1–10.

CODEBOOK

Codebooks are used by survey researchers to serve two main purposes: to provide a guide for coding responses and to serve as documentation of the layout and code definitions of a data file. Data files usually contain one line for each observation, such as a record or person (also called a "respondent"). Each column generally represents a single variable; however, one variable may span several columns. At the most basic level, a codebook describes the layout of the data in the data file and describes what the data codes mean. Codebooks are used to document the values associated with the answer options for a given survey question. Each answer category is given a unique numeric value, and these unique numeric values are then used by researchers in their analysis of the data.

As a guide for coding responses, a codebook details the question-and-answer wording and specifies how each individual answer should be coded. For example, a codebook entry for a question about the respondent's gender might specify that if "female" is chosen, it should be coded as "1," whereas "male" should be coded as "2." Directions may also be given for how to code open-ended answers into broad categories. These values are then used to enter the data the values represent into the data file, either via computer-assisted data entry software or in a spreadsheet.

There are many ways to create a codebook. Simple codebooks are often created from a word processing version of the survey instrument. More complex codebooks are created through statistical analysis software, such as SAS or Statistical Package for the Social Sciences (SPSS). Codebooks generated through statistical analysis software will often provide a variable label for each question, describing the content of the question, word and numeric labels for all answer categories, and basic frequencies for each question.

Codebooks can range from a very simple document to a very complex document. A simple codebook will detail each question-and-answer set along with the numeric value assigned to each answer choice, whereas a more complex codebook will also provide information on all associated skip patterns as well as any variables that have been "created" from answers to multiple other questions.

There are seven types of information that a codebook should contain. First, a short description of the study design, including the purpose of the study, the sponsor of the study, the name of the data collection organization, and the specific methodology used including mode of data collection, method of participant recruitment, and the length of the field period. Second, a codebook needs to clearly document all of the sampling information, including a description of the population, methods used to draw the sample, and any special conditions associated with the sample, such as groups that were oversampled. Third, the codebook needs to present information on the data file, including the number of cases and the record length of each case. Fourth, the data structure needs to be clearly delineated, including information on whether the data are presented in a hierarchical manner or some other manner. Fifth, specific details about the data need to be documented, including, at the very least, the variable names, the column location of each variable, whether the variable is numeric or character (string), and the format of numeric variables. Sixth, the question text and answer categories should be clearly documented along with frequencies of each response option. Finally, if the data have been weighted, a thorough description of the weighting processes should be included.

Major survey research projects conducted for the federal and state government often create electronic versions of codebooks that are accessible through the agencies' Web sites. There are also numerous centers and libraries at universities that provide archives of survey data from research projects along with Web access to electronic codebooks.

Lisa Carley-Baxter

See also Coder Variance; Coding; Frequency Distribution; Recoded Variable

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Coder Variance

Coder variance refers to nonsampling error that arises from inconsistencies in the ways established classification schemes are applied to the coding of research observations. In survey research, coder variance is associated with the process of translating the raw or verbatim data obtained from open-ended survey items into a quantitative format that can be analyzed by computers.

To appreciate how coder variance can occur, it is useful to review the process of preparing open-ended survey item data for analysis. Once all or a representative sample of the data have been collected, verbatim answers are examined for the purpose of defining a list of response categories (i.e., "code labels") that may be used for shorthand representations of the item data collected from each respondent. This list is known as the "coding frame" for the open-ended survey item. Depending on the coding protocol established, exactly one element or multiple elements of the coding frame may be associated with the item data.

Members of the research team designated as "coders" are entrusted with the responsibility of examining each verbatim response given to an open-ended item and assigning one or more of the elements of the coding frame to represent that data. Coders attempt to perform their task in such a manner that another coder would choose the identical set of elements from the coding frame. However, since judgment in interpreting both the raw verbatim data and the coding frame elements themselves is involved, inconsistency in the use of the coding frame elements (or code labels) is inevitable.

Any differences or inconsistencies in the combination of coding frame elements assigned to represent the actual verbatim data across interviewers constitute coder variance. These inconsistencies can arise as the consequence of four types of error:

1. Encoding error is introduced when the coding frame fails to feature code labels that are sufficiently exhaustive to clearly capture and discriminate the information in the verbatim data. Thus, when coders encounter data not well reflected in the coding frame, they must choose among imperfect alternatives. This promotes inconsistencies in the assigned code labels chosen across coders.
2. Interpretation error occurs when different coders haphazardly draw different meanings or nuances from the data. When this happens, different coders may apply different code labels from the coding frame to represent the data.
3. Coding error is a consequence of incorrect or inconsistent application of the code labels to the verbatim data. Because coding frame labels are highly condensed shorthand for highly varied, often detailed, and nuanced information, coders may interpret the meanings of these condensed labels in varied ways that, in turn, result in inconsistencies in their applications across coders.
4. Systematic coder bias arises from the tendencies of coders—human beings who possess personal

biases, either innate or learned—toward avoidance or overuse of specific elements in the coding frame.

Researchers examining the phenomenon of coder variance typically have found it to be a substantial problem for some survey items and a relatively inconsequential concern for others. When truly a problem, coder variance can account for as much as half of all nonsampling error in the statistical estimates produced for an item. Likewise, even when components of coder variance are small, the loss of precision in statistical estimates can be substantial. Indeed, coder variance can reduce the statistical reliability of survey estimates to a level achievable with half the sample size in the absence of coder variance.

While it is impossible to anticipate the extent of error that coder variance is likely to introduce into an item's results, studies have shown that the lion's share of the unreliability associated with coder variance results from the use of code labels that are general in nature or included as "catch-all" codes. Thus, researchers who choose to include open-ended survey questions should recognize the inherent unreliability and limited value of such items unless they (a) take pains to develop coding frames featuring only highly nuanced and specific code labels and (b) engage their coders in detailed training regarding the meaning and assignment of code labels.

Jonathan E. Brill

See also Coding; Element; Open-Ended Question; Variance; Verbatim Responses

Further Readings

Kalton, G., & Stowell, R. (1979). A study of coder variability. *Applied Statistics*, *28*(3), 276–283.

CODING

Coding is the procedural function of assigning concise and specific values (either alpha or numeric) to data elements collected through surveys or other forms of research so that these data may be quickly and easily counted or otherwise processed and subjected to statistical analyses, most often using a computer. These values may be alphanumeric in format, although it is common practice to use entirely numeric characters or entirely alphabetical characters when assigning labels. Numeric character values generally are almost universally referred to as "numeric codes" while alphabetical character values (and sometimes alphanumeric labels) are commonly referred to in several fashions, including "strings," "string codes," and "alpha codes," among others.

Inasmuch as data processing and analysis is typically accomplished through the use of specialized computer application software programs (e.g., Statistical Package for the Social Sciences [SPSS] or SAS), the assignment of designated values permits data to be transferred from the data collection instrument (which itself may be an electronic system, such as a computer-assisted telephone interviewing network) into a compact, computer-readable, database form.

The process of value development and specification may occur at any of several points in time during the conduct of the research project.

Precoding refers to code development and specification that occurs prior to the commencement of data collection activities. Precoding is appropriate for those data elements of the study where observations (e.g., respondent responses to survey questions) can be anticipated and exhaustively (or nearly exhaustively) specified before the research data are collected. As such, in survey research, precoding is routinely employed for all closed-ended items, all partly closed-ended items, and certain open-ended questions with which the investigator can anticipate the exhaustive range or set of possible responses. In addition, precoding occurs naturally and virtually automatically for open-ended items where clear constraints pertaining to the respondent's answer are implied by the question itself—for example, *How many times, if any, in the past year did you visit a dentist for any type of dental care?*—and, for this reason, such questions are said to be "self-coding."

In contrast, *postcoding* refers to code development and assignment that occur after data collection activities have begun. Most often, *postcoding* refers to code development and specification procedures implemented after the completion of data collection. However, to reduce the length of time between the data collection and subsequent data analysis activities of a study, postcoding might be initiated during data collection whenever a reliable subset of the full data set has been collected or when there is prior experience with similar questions.

Precoded labels are typically assigned in a manner that coincides with the measurement level implied by the item. For example, code labels assigned to response possibilities that correspond to interval or ratio level measures typically are numerical, with number values chosen to reflect the ordered and evenly spaced characteristics assumed by these measurement levels. (If a ratio level of measurement is involved, the code "0" is assigned to represent the measure's zero value.) Similarly, when ordinal level measurement items are involved, numerals (rather than alphabetical characters) are typically used for the codes, and the number values chosen appear in a logical sequence that is directionally consistent with the ordinal character of the measure's response categories; for example, 1 = None of the time, 2 = Some of the time, 3 = Most of the time, and 4 = All of the time. In contrast, code labels for items featuring nominal levels of measurement may be assigned in an arbitrary manner, as they bear no meaning or relationship to the response categories themselves; for example, 1 = No, 2 = Yes, or N = No, Y = Yes. Therefore, while sequenced numerals may be used for the code labels, these are typically assigned in an order corresponding to the sequence in which the response choices are documented in the research instrumentation. In other cases with nominal variables, simple alpha codes might be used, the convention often being using the first letter of the response choice.

Postcoding operations in survey research are bound to the categorization and structuring of responses culled from open-ended items, questions where the respondent's answers are self-composed and subject to unpredictable variation. To convert such data to computer-readable form, responses need to be associated with uniform categories and designated codes (typically numerals rather than letters) for these categories need to be assigned.

There are two approaches to accomplishing these postcoding tasks. One possibility is to develop a coding scheme prior to data collection activities. This approach requires that there is some theoretical basis for anticipating the possible responses and/or that the investigator has knowledge of and/or experience with a similar question or questions in one or more previous studies. The other possibility requires waiting until data collection activities have been completed or, alternately, until a representative subset (e.g., 20%) of the data have been collected. The available data are then examined for the purpose of establishing categories that capture the breadth and depth of the information collected and then assigning code labels to correspond to these categories. Then, once categories and corresponding labels have been established, item data for each interview are reviewed and one or more of these code labels are assigned to represent the information that was collected.

Standard research practice is to document the coded label values for each planned research observation (i.e., survey interview item) in a codebook. This document is more than just a listing of coded values, however; it is a blueprint for the layout of all information collected in a study. As such, the codebook not only identifies the value assigned to each research datum (i.e., survey answer, observation, or measurement) and the name of that value (i.e., the value label), but it also documents each label's meaning, specifies the name used to identify each item (i.e., "variable name"), includes a description of each item ("variable label"), and defines the data structure and reveals the specific location within that structure in which coded label values are stored.

Jonathan E. Brill

See also Closed-Ended Question; Codebook; Content Analysis; Interval Measure; Nominal Measure; Open-Ended Question; Ordinal Measure; Precoded Question; Ratio Measure; SAS; Statistical Package for the Social Sciences (SPSS)

Cognitive Aspects of Survey Methodology (CASM)

The cognitive aspects of survey methodology (CASM) is the interdisciplinary science involving the intersection of cognitive psychology and survey methods. CASM research endeavors to determine how mental information processing by respondents influences the survey response process and ultimately the quality of data obtained through self-report (or by proxy). CASM is mainly concerned with the study of response tendencies involving questionnaire data collection, but it can be more broadly defined as involving any aspect of survey-related mental processing, including respondent perceptions of survey interviewers and the survey introductions they use, the effects of administration mode (paper, telephone, computer), or responses to private or otherwise sensitive topics.

Background and History

Following the cognitive revolution of the 1970s, in which cognition was applied to a wide range of behavioral domains, the CASM field developed as an approach to questionnaire design that emphasizes the vital importance of cognition in the survey response process. Although the origins of this interdisciplinary science are rooted in earlier work, CASM as an identifiable movement was initiated by two key events: (1) the 1983 Advanced Research Seminar on Cognitive Aspects of Survey Methodology in the United States, now referred to as CASM I, and (2) the 1984 Conference on Social Information Processing and Survey Methodology held at ZUMA in Germany.

One influential outcome of the CASM I conference was the introduction of the four-stage cognitive model by Roger Tourangeau. To a great extent, the CASM approach is predicated on the key assertion that in order for a respondent to provide an accurate answer to a survey question, that individual must successfully negotiate a series of mental processing steps:

1. Comprehension of the survey question in the manner intended by the designer
2. Recall or retrieval from memory of information necessary to answer the question correctly
3. Decision and estimation processes that are influenced by factors such as item sensitivity, social desirability, or the respondent's assessment of the likelihood that the retrieved information is correct
4. The response process, in which the respondent produces an answer to the question in the form desired by the data collector

Some authors have elaborated this basic cognitive model by introducing other processes or mental states, such as motivational level. Others have envisioned a more flexible processing chain, in which the order of cognitive processes, and whether each is operative in a given case, varies depending on the survey question, the particular respondent, and the environment in which data collection occurs (e.g., the physical and social context).

Applied and Basic CASM Research

The CASM orientation has generated a wide range of research, which Monroe Sirken and colleagues have categorized as falling within two fundamental areas: applied CASM research and basic CASM research. Applied CASM research is focused on a specific questionnaire and attempts to improve that instrument through the use of cognitive interviewing methods to identify defects in survey questions having a cognitive origin. Basic CASM research is more general in scope. Rather than focusing on a particular instrument, basic CASM studies are devoted to the use of experimental methods to identify consistent cognitive tendencies that impact survey responding. Basic cognitive research is therefore intended to be applicable across a range of surveys and to serve as a guide to initial question design, rather than as a tailored pretesting method. That is, as opposed to focusing on quality control concerning a particular instrument, basic CASM research strives to elucidate rules of questionnaire design that incorporate a cognitive focus and that are developed through the use of empirical experimentation.

Examples of Basic CASM Research Studies

Some of this experimentation has concerned issues of response order effects, or how the respondent's tendency to select a particular response category (e.g., choice of a vague quantifier such as *excellent, very good, good, fair, poor*, or *very poor*) may depend on the order in which these options appear. Experiments by Jon Krosnick and colleagues have determined that response order effects depend on factors such as survey administration mode, for reasons having a cognitive basis. When response categories appear visually, as on a self-administered instrument, a *primacy effect* is often observed, where respondents are more likely to select items early in the list, presumably due to motivational factors such as satisficing that lead to fuller processing of earlier items than later ones. On the other hand, when the same response categories are read aloud under interviewer administration, a *recency effect* is obtained, in which later items in the list are more likely to be selected. From a cognitive point of view, recency effects are hypothesized to occur due to short-term memory limitations, where the items read most recently (those later in the list) are better represented in the respondent's memory and are therefore favored.

As a further example of experimentally oriented basic CASM research, Norbert Schwarz and colleagues cited in Tourangeau et al. have considered the effects of open-ended versus closed response categories for questions that ask about the frequency and duration of common, mundane behaviors. Their results suggest that respondents make use of information that is implicitly conveyed through such design decisions. In one experiment, subjects were asked to estimate the number of hours per day that they watched television, but one group was given closed-ended response categories ranging between "Up to ½ hour" through "More than 2½ hours" (low range), and the other was presented ranges from "Up to 2½ hours" through "More than 4½ hours" (high range). Individuals in the low-range condition tended to select a relatively lower duration of television watching than did those presented the higher ranges. The investigators concluded that respondents in both situations considered the middle category to represent normative or expected behavior and therefore relied on this central value as an anchor point when selecting their own answer from the presented list. Given the potentially contaminating effect of such response category ranges, the investigators suggested that designers instead choose an open-ended format for questions asking about behaviors like television watching, as this will obtain the desired information without subtly promoting any particular response category.

Similarly, CASM theorizing and research have concerned the effects of a number of other questionnaire design variables, such as (a) question ordering and its relationship to context effects, due to comprehension, memory, and decision-related processes; (b) variation in item sensitivity or degree of threat to personal privacy, which may influence respondents' decision making concerning the likelihood of providing a truthful response; (c) question length and complexity, which may affect overall cognitive processing burden; and (d) the effects of varying reference periods for recall of information, especially as this produces forward and backward telescoping effects.

Practical Use of Basic CASM Research Results

Basic CASM studies have been compiled and summarized in books by Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski and by Seymour Sudman, Norman Bradburn, and Norbert Schwarz. Questionnaire designers can rely on this body of evidence to determine the cognitive factors that are likely to influence responses to their questions and to consider design alterations expected to improve overall response quality (e.g., the use of an administration mode that removes the presence of a human interviewer when sensitive questions are asked). This body of evidence is certainly useful in providing guidance, as it considers vital design issues and is dependent on the results of controlled experimentation. An important limitation, however, is that such experimental results are often insufficient, in themselves, for purposes of directing design decisions in specific cases, because the "rules" that emanate from such results tend to be somewhat generic in nature and subject to exception. For example, the knowledge that longer questions generally tend to reduce comprehension, relative to shorter ones, will not reveal the optimal length for a particular combination of respondent population and survey topic. For this reason, the basic CASM research approach is supplemented by empirical pretesting techniques, such as cognitive interviewing and behavior coding, which represent the applied CASM orientation.

Extension to the General Study of Cognition

CASM research is intended by its proponents to ultimately forge a path toward a two-way street in which research findings benefit not only survey researchers, but as well inform the science of cognitive psychology. This outcome may be facilitated in part because the study of cognition within the survey context provides an environment that widens the scope of inquiry to naturalistic circumstances beyond those investigated within the typical psychological laboratory situations (e.g., memory for real-world autobiographical events). Further, CASM studies often involve a broad range of the population, in terms of demographic characteristics such as age and educational level, rather than focusing on college students as study subjects. Despite these potential benefits, however, the impact of CASM on the general field of cognitive psychology has to date been somewhat limited. Expanding this direction remains an endeavor that is ripe for further development.

Gordon B. Willis

See also Behavior Coding; Cognitive Interviewing; Context Effect; Primacy Effect; Recency Effect; Satisficing; Telescoping

Further Readings

Jabine, T. B., Straf, M. L., Tanur, J. M., & Tourangeau, R. (Eds.). (1984). *Cognitive aspects of survey methodology:*

Building a bridge between disciplines. Washington, DC: National Academy Press.

Jobe, J. B., & Mingay, D. J. (1991). Cognition and survey measurement: History and overview. *Applied Cognitive Psychology*, *5*, 175–192.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Schaeffer, N. C. (1999). Asking questions about threatening topics: A selective overview. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 105–122). Mahwah, NJ: Lawrence Erlbaum.

Sirken, M., Herrmann, D., Schechter, S., Schwarz, N., Tanur, J., & Tourangeau, R. (Eds.). (1999). *Cognition and survey research.* New York: Wiley.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Cognitive Interviewing

Cognitive interviewing is a psychologically oriented method for empirically studying the ways in which individuals mentally process and respond to survey questionnaires. Cognitive interviews can be conducted for the general purpose of enhancing the understanding of how respondents carry out the task of answering survey questions. However, the technique is more commonly conducted in an applied sense, for the purpose of pretesting questions and determining how they should be modified, prior to survey fielding, to make them more understandable or otherwise easier to answer.

The notion that survey questions require thought on the part of respondents is not new and has long been a central premise of questionnaire design. However, cognitive interviewing formalizes this process, as it approaches the survey response task from the vantage point of cognition and survey methodology (CASM), an interdisciplinary association of cognitive psychologists and survey methodologists. The cognitive interview is generally designed to elucidate four key cognitive processes or stages: (1) comprehension of the survey question; (2) retrieval from memory of information necessary to answer the question; (3) decision or estimation processes, especially relating to the adequacy of the answer or the potential threat it may pose due to sensitive content or demands of social desirability; and (4) the response process, in which the respondent produces an answer that satisfies the task requirements (e.g., matching an internally generated response to one of a number of qualitative response categories on the questionnaire).

For example, answering the survey question *In the past week, on how many days did you do any work for pay?* requires that the respondent comprehends the key elements "week" and "work for pay," as well as the overall intent of the item. He or she must retrieve relevant memories concerning working and then make a judgment concerning that response (for instance, the individual may have been home sick all week, but in keeping with the desire to express the notion that he or she is normally employed, reports usual work status). Finally, in producing a response, the respondent will provide an answer that may or may not satisfy the requirements of the data collector (e.g., "Four"; "Every day"; "Yes, I worked last week"). The cognitive model proposes that survey questions may exhibit features that preclude successful cognitive processing and that may result in survey response error (in effect, answers that are incorrect). In the preceding example, the question may contain vague elements ("week"; "work for pay") that create divergent interpretations across respondents; or it may induce biased responding (e.g., the socially desirable impulse to provide a nonzero response).

Cognitive Interviewing Procedures

The major objective of cognitive interviewing is to identify sources of response error across a wide range of survey questions, whether autobiographical (involving behavior and events), attitudinal (involving opinions and attitudes), or knowledge based. To this end, a specially trained cognitive interviewer administers the questions individually to persons (often referred to as "laboratory subjects") who are specifically recruited for purposes of questionnaire evaluation or pretesting. In departure from the usual question-and-answer sequence within a survey interview, the cognitive interview involves procedures designed to delve into the cognitive processes that underlie the production of the answers to evaluated questions, by inducing the subject to produce verbal reports.

Two related procedures are used to elicit verbal reports: *think aloud* and *verbal probing*. The think-aloud procedure was adapted from psychological laboratory experiments and requires subjects to verbalize their thoughts as they answer survey questions. The interviewer prompts the subject as necessary by providing feedback such as "Tell me what you are thinking" or "Keep talking." The researchers then analyze the resulting verbatim verbal stream to identify problems in answering the evaluated questions that have a cognitive origin. For example, the subject's verbal protocol relating to the preceding question on work status might include a segment stating, "Besides my regular job, last Saturday I, uh, did help a friend of a friend move into a new apartment—he gave me pizza and beer—and a gift card that was lying around with a little money on it still, so I guess you could call that working for pay, but I'm not sure if that's supposed to count." Given this accounting, the investigators might surmise that the meaning of "work for pay" is unclear, in this case concerning irregular work activities that result in noncash remuneration. Especially if this finding were replicated across multiple cognitive interviews, the questionnaire designer could consider revising the question to more clearly specify the types of activities to be included or excluded.

However, practitioners have observed that some subjects are unable to think aloud effectively, and that the pure think-aloud approach can be inefficient for purposes of testing survey questions. Therefore, an alternative procedure, labeled "verbal probing," has increasingly come into prominence and either supplements or supplants think aloud. Probing puts relatively more impetus on the interviewer to shape the verbal report and involves the use of targeted probe questions that investigate specific aspects of subjects' processing of the evaluated questions. As one common approach, immediately after the subject answers the tested question, the interviewer asks probes such as "Tell me more about that"; and "What does the term 'work for pay' make you think of?" Probe questions are sometimes designed to tap a specific cognitive process (e.g., comprehension probes assess understanding of the question and its key terms; retrieval probes assess memory processes). However, probes also lead the subject to provide further elaboration and clarify whether the answer provided to the evaluated question is consistent with and supported by a picture gleaned through a more thorough examination of the subject's situation.

Verbal probing can be used to search for problems, proactively, when probes are designed prior to the interview, based on the anticipation of particular problems. Or, probes may be reactive, when they are unplanned and are elicited based on some indication by the subject that he or she has some problem answering it as intended (e.g., a delay in answering or a response that seems to contradict a previous answer). The proactive variety of probing allows the cognitive interviewer to search for covert problems that otherwise do not surface as a result of the normal interchange between interviewer and subject. Conversely, reactive probes enable follow-up of unanticipated overt problems that emerge.

Further, the type of probing that is conducted depends fundamentally on variables such as survey administration mode. For interviewer-administered questions (telephone or in person), probes are often administered concurrently, or during the conduct of the interview, immediately after the subject has answered each tested question. For self-administered questionnaires in particular, researchers sometimes make use of retrospective probes, or those administered in a debriefing step after the main questionnaire has been completed, and that direct the subject to reflect on the questions asked earlier. Concurrent probing provides the advantage of eliciting a verbal report very close to the time the subject answers the tested questions, when relevant information is likely to remain in memory. The retrospective approach risks the loss of such memories due to the delay between answering the question and the follow-up probes. On the other hand, it more closely mirrors the nature of the presentation of the targeted questions during a field interview (i.e., uninterrupted by probes) and prompts the subject to reflect over the entire questionnaire. Cognitive interviewing approaches are flexible, and researchers often rely both on concurrent and retrospective probing, depending on the nature of the evaluated questionnaire.

Analysis of Interview Results

Concerning analysis of obtained data, the focus of cognitive interviewing is not primarily the answers to tested questions, or quantitative data, but rather qualitative data relevant to the evaluation of tested questions. Cognitive interviews normally produce data in the form of written notes taken by the interviewer during the course of the interview, of notes taken by observers, or of analysis of (audio or video) recordings. Such analyses sometimes depend on a coding scheme that applies a particular category of outcome to subjects' behaviors or to interviewer comments (e.g., identification of a "vague term"). More often, however, data derived from cognitive interviews consist of written summaries that describe the problems observed on a question-by-question basis, across a set of interviews, and that also propose modifications intended to address these problems. On the basis of these results and suggestions, the investigators may revise the questions and then conduct further sets, or *rounds*, of cognitive testing. Such iterative testing rounds are useful for determining if the proposed solutions have solved identified problems without introducing additional difficulties.

Logistics of Cognitive Interviewing

Because the major emphasis of the cognitive interview is not survey data collection but rather the efficient and timely development and evaluation of survey questions in an applied setting, sample sizes for a round of cognitive interviews are generally small; typically between 8 and 12 subjects. In departure from the random selection procedures of the field survey, cognitive interviewing most often depends on volunteers who are recruited explicitly to represent as wide as possible a range of the population to be surveyed, primarily through the use of newspaper advertisements and posted flyers, or visits by researchers to locations where eligible individuals can be located (e.g., a clinic, service agency, school, or elderly center). Cognitive interviews are often conducted within permanent questionnaire design laboratories staffed by trained and experienced professionals and recruitment specialists, but they can also be accomplished informally by a questionnaire designer for the purpose of evaluating a single questionnaire. Within a laboratory environment, cognitive interviewing is conducted as one component of a more comprehensive pretesting process that includes additional pretesting procedures such as review by subject matter experts and focus groups (which normally precede cognitive interviews), or behavior coding (which is generally conducted after cognitive interviewing rounds, as part of a survey field pretest).

Variation in Practice

Although cognitive interviewing is a common and well-established pretesting and evaluation method, the precise activities that are implemented by its practitioners vary in key respects. Cognitive testing of questionnaires used in surveys of businesses and other establishments places significant emphasis on information storage and retrieval, especially because relevant information is often retained in administrative records rather than respondent memories and is distributed among multiple sources. For any type of survey, questions that focus on sensitive information (e.g., drug use, sexual behavior, or income) tend to focus on decision processes that influence the truthfulness of responses.

Practitioners also vary widely with respect to how they conduct the interviews, concerning reliance on think aloud versus verbal probing, and whether the cognitive interviews are conducted by researchers who will also serve as analysts or by an interviewing team that will present the testing results to the investigators for further consideration. At this time it is not clear which of these approaches are most reliable or valid, although researchers have recently begun rigorously to evaluate the effectiveness of cognitive interviews in various guises.

Researchers have recently focused increasingly on cultural as well as cognitive aspects of survey questions. One promising new direction, therefore, is the use of the cognitive interview to assess the cross-cultural comparability of questions, especially when they are translated from a source language into one or more target languages. As such, cognitive interviewing procedures are extended to diverse population subgroups to determine whether these questions function appropriately across group or language. Further, although cognitive interviewing has mainly been applied to survey questionnaires, practitioners have also begun to use this method to assess a wide range of other survey-relevant materials, such as advance letters to survey respondents, survey introductions used by interviewers to gain respondent cooperation, research consent forms, statistical maps and graphs, and computer Web sites (in a manner very similar to usability testing). The cognitive interview is in principle applicable in any case in which researchers wish to investigate the ways in which individuals understand and react to orally or visually presented materials that demand mental processing activity.

Gordon B. Willis

See also Behavior Coding; Cognitive Aspects of Survey Methodology (CASM); Focus Group; Language Translations; Pretest; Usability Testing

Further Readings

Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, J. Rothgeb, M. Couper, J. Lessler,

E. Martin, J. Martin, et al. (Eds.), *Questionnaire development evaluation and testing methods* (pp. 45–66). Hoboken, NJ: Wiley.
Conrad, F., & Blair, J. (2004). Data quality in cognitive interviews: The case for verbal reports. In S. Presser et al. (Eds.), *Questionnaire development evaluation and testing methods* (pp. 67–87). Hoboken, NJ: Wiley.
DeMaio, T. J., & Rothgeb, J. M. (1996). Cognitive interviewing techniques: In the lab and in the field. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 175–195). San Francisco: Jossey-Bass.
Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.
Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 393–418). New York: Wiley.
Willis, G. B. (1999). *Cognitive interviewing: A how-to guide.* Retrieved March 24, 2008, from http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf
Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Cold Call

A cold call refers to the circumstance that takes place in many surveys when a respondent is first called or contacted in person by a survey interviewer without any advance knowledge that he or she has been sampled to participate in the survey, and thus does not know that the call or contact is coming. This circumstance contrasts to other instances in which some form of advance contact has been made with the sampled respondent to alert him or her—that is, to "warm up" him or her—that he or she has been sampled and that an interviewer soon will be in contact. Survey response rates consistently have been found to be lower for those sampled respondents that receive cold calls than for those that receive advance contact.

For many people who are sampled in telephone surveys, there is no way that researchers can use an advance mail contact technique because all that is known about the sampled household is the telephone number. This occurs even after the researchers have run matches of sampled telephone numbers against address databases and no address match is identified. Granted, an advance telephone contact attempt could be made in which a recorded message is left alerting the respondent that he or she has been sampled for a survey and that an interviewer will call him or her within a few days. However, there is no reliable evidence that this approach ever has been found to be effective. Instead the concern is that such a telephonic advance contact will lower response propensity at the given telephone number when the human interviewer eventually makes contact.

Despite this concern, the argument can be made that advance telephone contacts that merely leave a recorded message that a household has been chosen for a survey are not dissimilar to instances in which interviewers reach an answering machine the first time they call a household and leave a message saying that they will be calling back to conduct a survey. Past research has found that these types of answering machine messages tend to raise response rates. As such, even with households that cannot be mailed an advance contact, the proportion that receives cold calls for telephone surveys can be greatly reduced.

With face-to-face interviewing in address-based sampling or area probability sampling, all sampled households can be mailed an advance contact because, by definition, the researchers know their addresses. Thus, in such surveys there are no structural barriers that make it impossible to avoid any household receiving a cold contact from the in-person interviewer when he or she arrives the first time to recruit the household and/or gather data.

Paul J. Lavrakas

See also Advance Contact

Further Readings

de Leeuw, E., Callegaro, M., Hox, J. Korendijk, E., & Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys: A meta-analysis. *Public Opinion Quarterly*, *71*(3), 413–443.

Common Rule

The Common Rule refers to a set of legal and ethical guidelines designed for protection of human subjects in research either funded by federal agencies or taking place in entities that receive federal research funding. The term *Common Rule* technically refers to all the regulations contained in Subpart A of Title 45 of the Code of Federal Regulations Part 46 (45 CFR 46). As applied to survey research, the most important elements of the

Common Rule are those relating to oversight by an institutional review board and the requirements of informed consent and voluntary participation.

Background

In the early 1970s, a number of high-profile cases of clearly unethical research made headlines and resulted in calls for congressional hearings. A few of the most striking examples include the following:

- *The Tuskegee Syphilis Study (1932–1972)*. Begun in 1932 to test syphilis treatments, the federal Public Health Service enrolled hundreds of African American men to participate. Deception was a key feature of the research from the start, but it was taken to new levels in the 1940s, after penicillin was proven an effective cure for syphilis. The researchers prevented their subjects from obtaining beneficial medical treatment and maintained their deception until 1972, when details of the study first came out in the press. The study directly caused 28 deaths, 100 cases of disability, and 19 cases of congenital syphilis and was in direct violation of several elements of the Nuremberg Code (1945), developed after World War II in response to Dr. Joseph Mengele's infamous experiments on Nazi concentration camp victims.
- *Milgram's Experiments on Obedience to Authority*. In attempting to determine the extent to which typical Americans might be willing to harm others simply because an authority figure told them to, psychologist Stanley Milgram designed an experiment in the early 1960s in which the subjects believed that they were delivering ever-stronger electrical shocks to a "learner" who was actually part of the research team. A large majority of subjects continued to comply even after they believed they were causing severe pain, unconsciousness, and even, potentially, death. Very early on, subjects showed clear signs of severe psychological stress, but Milgram continued his experiments to the end, even adding an especially cruel treatment condition in which the subject had to physically hold the "victim's" hand in place. (The ethics of Milgram's work has been debated for years, but many believe that it served a very positive role in showing the power and danger of authoritarianism and also served as an important warning to the scientific community for the need to make more formal and stringent ethical procedures for all social research.)
- *Zimbardo's Prison Experiment*. As part of a research study, and after randomly assigning student volunteers to be either "prisoners" or "guards" in the early 1970s, psychologist Philip Zimbardo found that members of both groups were taking to their roles to a much greater extent than he had anticipated. Despite clear indications within 36 hours that some of the students were deeply stressed by participating in the study, the experiment was continued for 6 full days.

The Milgram and Zimbardo experiments, in particular, served as wake-up calls to social science researchers who, until that point, had generally considered research ethics a topic of interest to medical research but not to the social sciences. In both cases the unethical behavior occurred not so much with regard to the research designs but rather with regard to the choices the researchers made after their studies went in unanticipated harmful directions. The principal investigators decided to continue their experiments long after they were aware of the harm they were causing their research subjects, a fact that made comparisons to the Tuskegee Experiment both inevitable and appropriate. Indeed, by failing to balance the anticipated benefits of the research with the risks to their subjects, they were in violation of a key provision of the Nuremberg Code.

Congressional and Regulatory Action

As a result of press reports and resultant public outcries about these cases, Congress held hearings in 1973 titled "Quality of Health Care—Human Experimentation." The hearings led to the passage of the National Research Act of 1974, which established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research and required the creation of institutional review boards (IRBs) at all institutions receiving funding from the Department of Health, Education, and Welfare (HEW).

The commission was charged "to identify the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and to develop guidelines . . . to assure that such research is conducted in accordance with those principles." The first regulations were issued as 45 CFR 46, "Regulations for the Protection of Human Subjects of Biomedical and Behavioral Research," in 1974 by HEW (now Health and Human Services, or HHS); these were revised and expanded on after the release of the commission's report in April 1979. The Belmont Report first laid out three "Basic Ethical Principles": (1) respect for persons, (2) beneficence, and (3) justice. Then it detailed specific ways in which those principles should be applied in practice and focused especially on the importance of informed consent, assessment of risk and benefits, and the selection of subjects. These provisions of the Belmont Report are now encoded in 45 CFR 46 section 111, leading some researchers to use the terms *Belmont Report* and *Common Rule* interchangeably.

After revisions to the regulations in 1991, 16 other federal agencies adopted them, leading to their current informal name, the Common Rule. Thus, the provision requiring all institutions that receive federal research funds to establish IRBs now includes federal funds from virtually any federal agency. As a result, virtually all colleges and universities now have IRBs.

Applicability to Survey Research

According to subsection 101 of the regulations, survey research is not subject to IRB review unless "human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation." Nonetheless, most university IRBs still require at least expedited review of survey research conducted under their auspices to ensure that the basic principles outlined in the Belmont Report and encoded in the Common Rule are observed.

Although survey research only rarely poses the sorts of ethical dilemmas or risks to human subjects found in medical research, or even psychological experimentation, many survey researchers consider it a matter of best practices to abide by most elements of the Common Rule. For example, although even survey research projects conducted under the supervision of university IRBs generally are not required to undergo the full process of informed consent, they generally are required to assure respondents of the confidentiality and/or anonymity of their responses and the voluntary nature of their participation. In fact, this norm is so strong that most non-academic survey researchers include some form of these assurances even though they are not covered by an IRB or by legal regulations.

IRBs provide especially strong oversight over surveys that focus on sensitive topics that might place respondents under stress. These areas would include drug and alcohol use, criminal behavior, sexual behavior, and experiences of victimization or discrimination. In addition, surveys of vulnerable populations—minors, mentally or developmentally disabled adults, and prison inmates—are also generally subject to a higher level of oversight.

But even when conducting research that is not covered by IRB oversight or that does not meet any legal definitions that would seem to require special attention to the rights of human subjects, survey researchers would do well to keep in mind the principles of the Common Rule. Survey response rates have already declined a great deal due to growing public resistance to survey research among the general public, fed by a variety of deceptive tactics such as push polls and FRUGing (fund-raising under the guise of survey research). In this environment, attention by legitimate survey researchers to the basic ethical principles of respect for persons, beneficence, and justice will be crucial to ensuring the viability of survey research in the future.

Joel D. Bloom

See also Anonymity; Beneficence; Common Rule; Confidentiality; Deception; Ethical Principles; FRUGing; Informed Consent; Institutional Review Board; Minimal Risk; Protection of Human Subjects; Push Polls; Survey Ethics; Voluntary Participation

Further Readings

Citro, C., Ilgen, D., & Marrett, C. (2003). *Protecting participants and facilitating social and behavioral sciences research.* Washington, DC: National Academy Press.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* New York: Wiley.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* Retrieved March 24, 2008, from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm

U.S. Department of Health and Human Services. (2005). *Code of Federal Regulations, Title 45 Public Welfare and Part 46 Protection of Human Subjects.* Retrieved March 17, 2008, from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

COMPLETED INTERVIEW

The completed interview survey disposition is used in all types of surveys, regardless of mode. In a telephone or in-person interview, a completed interview results when the respondent has provided answers for all of the questions on the survey questionnaire that were asked by the interviewer. In a mail survey, a completed interview results when the respondent receives a paper-and-pencil survey questionnaire, answers all questions on the questionnaire, and returns the completed questionnaire to the researcher. In an Internet survey, a completed interview occurs when the respondent logs into the survey, enters answers for all of the questions in the questionnaire, and submits the questionnaire electronically to the researcher. Completed interviews are eligible cases and are considered a final survey disposition.

It is worthwhile to note that a completed interview usually indicates that the respondent has provided data (answers) for all applicable items on a questionnaire. However, at times respondents may answer most of the questions on a questionnaire but may accidentally skip or refuse to answer some questions on the survey instrument (called "item nonresponse"). Depending on how much data are missing, these interviews may be considered partial completions due to this item nonresponse but may also be considered breakoffs (or refusals) if the respondent began the interview or questionnaire but answered only a few of the applicable questions.

In practice, the level of item nonresponse may be very small, and it may be difficult to differentiate a completed interview from a partial interview. For this reason, most survey organizations have developed rules that explicitly define the differences among breakoffs, partial interviews, and completed interviews. Common rules used by survey organizations to determine whether an interview with item nonresponse can be considered a completed interview include (a) the proportion of all applicable questions answered; and (b) the proportion of critically important or essential questions administered. For example, cases in which a respondent has answered fewer than 50% of the applicable questions might be defined as breakoffs; cases in which the respondent has answered between 50% and 94% of the applicable questions might be defined as partial completions; and cases in which the respondent has answered more than 94% of applicable questions might be considered completed interviews.

Matthew Courser

See also Final Dispositions; Missing Data; Partial Completion; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Completion Rate

The term *completion rate* has been used often in the survey research literature to describe the extent of cooperation with and participation in a survey. However, it is an ambiguous term because it is not used consistently. Therefore readers of the literature should interpret the term with caution.

Completion rate is often used to describe the portion of a questionnaire that has been completed. In self-administered surveys, it is used widely to differentiate between the number of eligible individuals who do not complete a questionnaire and those who do. In this context, the completion rate is the number of questionnaires completed divided by all eligible and initially cooperating sample members. Researchers using completion rate in this sense should state so explicitly. This rate is an important indicator of item nonresponse in self-administered surveys. It has implications for the visual layout of a self-administered instrument, since the layout may affect how willing sample members are to complete the questionnaire. In addition, it also has implications for the content and the placement of critical questions in the questionnaire.

Completion rate is also an umbrella term used to describe the extent of sample participation in a survey—including the response rate, the contact rate, and the cooperation rate. Since these outcome rates are often used as criteria for evaluating the quality of survey data, analysts and other data users should know which rate is being referred to by the term completion rate. The *response rate* indicates the proportion of the total eligible sample that participates in the survey, the *contact rate* indicates the proportion of those contacted out of all eligible sample members, and the *cooperation rate* indicates the proportion of the contacted sample that participates in (or consents to participate in) the survey. The American Association for Public Opinion Research (AAPOR) recommends that researchers define how they are using the terms *response rate, contact* *rate*, and *cooperation rate* and offers standard definitions for these terms and how they should be calculated. AAPOR recommends that researchers explain in detail how they calculated the rates and how they categorized the disposition codes. Of note, AAPOR does not define the calculation of the term *completion rate*.

In addition to responding to a survey, people may participate in studies in other ways as well, and instruments other than questionnaires are often used to collect data. For instance, a screener interview may be used to determine an individual's eligibility for a study before he or she is asked to participate in the full survey. In addition to self-reported information collected during an interview, other data may be collected from participants, such as biomeasure data (height and weight measures, hair samples, or saliva samples). In epidemiological or randomized controlled studies, sample members may be asked to participate in a health regimen, in special education programs, or in an employment development program. The term *completion rate* may therefore be used to indicate the extent to which any or all of these activities have been completed. This more or less "universal" nature of the term underscores the importance of defining how it is being used in any given context. For example, in reporting findings based on biomeasure data, researchers should be clear about whether *completion* means completing the questionnaire only or if they are referring to completing the additional data collection.

Because it is impossible to assign a term to every possible permutation of a survey, it is critical for researchers to fully explain the sense in which they are using terms such as completion rate. It is equally important to use the terminology defined by the standard-setting organization(s) in a given discipline so as to promote a common understanding and use of terms.

Danna Basson

See also Completed Interview; Cooperation Rate; Final Dispositions; Partial Completion; Response Rates; Standard Definitions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly*, *62*, 378–397.

Complex Sample Surveys

Complex sample surveys involve the identification and data collection of a sample of population units via multiple stages or phases of identification and selection. In contrast, a simple sample survey design involves a simple random sample, where there is a list of the elements of the population and a certain number of these elements is selected by drawing one at a time. The classic textbook example is when each element of the frame is numbered from 1 to N (i.e., population size) and then n (i.e., sample size) elements are drawn using a table of random numbers. By contrast, complex sample surveys may rely on stratification, clustering, multi-stage or multi-phase designs, unequal probability sampling, or multi-frame sampling. These techniques often reduce the cost of data collection and may be more efficient, but they also require special methods of variance estimation and in many cases yield larger variances than a simple random sample of the same size. Ultimately the objective of a complex sample design is to minimize variance and costs for all the desired estimates while preserving the ability to obtain valid point and variance estimates for population parameters of interest.

Stratification

One aspect of a complex sampling design may involve *stratification*, defined as a partition of the population into mutually exclusive and collectively exhaustive subsets called "strata." One primary reason for using stratification is usually associated with the recognition that members of the same stratum are likely to be more similar to each other than members of different strata. Other reasons for using stratification include the desire to have every part of the population represented, or the desire to reduce sampling variance by using a larger sampling fraction in strata when the unit variance is larger than in more homogeneous strata, or it may reflect a strategy based on differential data collection costs from stratum to stratum. Stratification could also be used if stratum-specific domain estimates are desired. As previously alluded to, the sampling fractions used within the different strata may or may not be the same across all the strata. Strata may be explicit, and the number of units to be selected from each strata may be determined beforehand. Or stratification may be implicit, when systematic sampling is used and the units are arranged with all the units in each stratum appearing together when the population is ordered. In the case where strata are explicit, algorithms such as Neyman allocations for single estimands or the Chromy allocation algorithm for multiple estimands may be used to decide how many units to select from each stratum. A minimum of 2 units per stratum is usually recommended, as this facilitates variance estimation.

Cluster Designs

While stratification attempts to partition the population into sets that are as similar to each other as possible, *clustering* tries to partition the population into sets that are as heterogeneous as possible, but where data collection is less expensive by selecting a number of clusters that contain population units. One example is in a survey of students in which a given number of schools are selected, and then students are sampled within each of those chosen schools or clusters. In this case, the schools are called the "primary sampling units" (PSUs), while the students within the schools are referred to as the "secondary sampling units" (SSUs). It is possible to take either a sample or census of the secondary sampling units contained within each of the selected clusters. This would be the case when sampling additional units is extremely inexpensive, such as sampling entire classrooms from selected schools. More common, however, is to select clusters as a first sampling stage and then to select a subset of units within the clusters as a second stage. Sometimes there are more than two stages within a design, such as when school districts are selected first, then schools within the districts, and then intact classrooms within the schools.

Another variant of cluster design is the multi-phase design. In this instance the clusters are selected as in a multi-stage design, but instead of selecting units within each cluster, units are selected from the union of all units within the selected clusters. Of course, depending on the assigned probabilities and selection method, some multi-phase designs are strictly equivalent to multi-stage designs.

Unequal Probability Designs

Whether a sampling design is stratified, clustered, or selected without any partitions of the population, one may select units with the same probability or with unequal probabilities. Or one may select PSUs with unequal probabilities in order to have each element of the population have the same probability of selection. Often the probability of selection is chosen to be proportional to some measure of size (i.e., sampling with probabilities proportional to size or PPS), particularly when sampling PSUs in a multi-stage or multi-phase sample. In order to achieve equal probabilities for each unit of the population, in a multi-stage design it is desirable to designate a probability of selection for every cluster that is proportional to the number of population units in the cluster and then to sample an equal number of units at the second stage. As with simple random sampling, the selection of clusters can be with or without replacement. A third option is to *sample with minimum replacement*, a term introduced by Chromy in 1979. According to such a design, the large PSUs (those that are to be sampled with certainty) may be sampled more than once. A decision to include PSUs multiple times in the final sample will usually depend on the intraclass correlation (rho)—a measure of how homogeneous are the clusters (PSUs).

Unequal probabilities may actually be used directly for the elements of the population and not just for the PSUs. One example is in an establishment survey by which one wants to determine the price of a particular product. If in an establishment survey the volume of sales of the product is listed for every element in the frame and one samples with PPS, when the volume of sales is the measure of size, a simple average of the prices charged by the establishments in the sample would yield an (unbiased) estimate of the average price of the units sold.

On the other hand, sometimes unequal probabilities may be used because there is a desire to oversample certain subpopulations. And sometimes a probability is calculated based on the need to obtain multiple estimates. For example, in an establishment survey in which the prices of different items need to be estimated and the volumes vary by the items, Chromy's allocation algorithm may be used to obtain a probability of selection for every establishment in the frame, but this probability of selection will not be proportional to any particular measure of size.

Weighting

The purpose of sampling in a survey is to obtain an estimate of a parameter in the population from which the sample was drawn. In order to do this, one must know how to weight the sampled units. The most common approach to weighting is to calculate a probability of selection and then take its multiplicative inverse. This yields the Horvitz-Thompson estimator, and though it seems straightforward, there are many designs for which this estimator is difficult or impossible to obtain. Dual-frame estimators represent a case in which the straightforward Horvitz-Thompson estimators have to be modified to incorporate the probability of being included into the sample via multiple frames. It is often the case that the initial weights (i.e., inverse of selection probability) are not the final versions used to produce the final estimates. Rather, the weights are often adjusted further to account for population sizes and/or nonresponse using a variety of techniques, including post-stratification, trimming of the weights, and the use of ratio or regression estimators.

Variance Estimation

Survey weights as well as the design upon which the weights are computed play an important role in both the parameter estimates and variance computations. Whereas estimating the variance of simple survey estimates is rather straightforward, variance estimation in complex sample surveys is much more complicated. Some sampling approaches have variance formulas that may be applied, but a multi-stage approach in which clusters are sampled with PPS and weight adjustments are made can be far more complex. There are two basic sets of methods that may be used: (1) Taylor series linearization and (2) replicate methods. In each of these methods it is important, although not always obvious, that the design be properly specified. One important consideration is that if a PSU is sampled with certainty, it must be treated as a stratum, and the units at the next level of sampling should be treated as PSUs.

Taylor series linearization has the advantage of using a straightforward approach that is available in many standard statistical packages. Replicate methods, such as the jackknife and balanced half sample pseudo-replications, allow one to reproduce aspects of the design, taking imputation into account. These methods are also available in many packages, but it is also easy to fail to specify the design properly. A more complex method is the bootstrap, which needs to be programmed specific to each design but allows for a closer reproduction of the initial sample.

Pedro Saavedra

See also Clustering; Multi-Stage Sample; *n*; *N*; Post-Stratification; Probability of Selection; Replicate Methods for Variance Estimation; ρ (Rho); Simple Random Sample; Stratified Sampling; Taylor Series Linearization; Variance Estimation; Weighting

Further Readings

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.

Composite Estimation

Composite estimation is a statistical estimation procedure that combines data from several sources, for example, from different surveys or databases or from different periods of time in the same longitudinal survey. It is difficult to describe the method in general, as there is no limit to the ways one might combine data when various useful sources are available. Composite estimation can be used when a survey is conducted using a rotating panel design with the goal of producing population estimates for each point or many points in time. If the design incorporates rotating groups, composite estimation can often reduce the variance estimates of level variables (e.g., totals, means, proportions). In addition, composite estimation can reduce the variance estimates of variables dealing with changes over time, depending on the structure of the sample design, the strength of the correlations between group estimates over time, and other factors.

How a Composite Estimator Works

In a typical rotation design, the sampled groups are phased in and out of the sample in a regular, defined pattern over time. To estimate the level of a characteristic in the time period designated by t, a simple compositing strategy is to take a convex combination of the Horvitz-Thompson estimate of level for period t, Y_t^{HT1}, with a second estimate for period t, Y_t^{HT2}. The latter estimate might start with the composite estimate for period $t-1$, Y_{t-1}^{CE}, brought forward by a measure of change from period $t-1$ to period t:

$$Y_t^{HT2} = Y_{t-1}^{CE} + \Delta_{t-1,t}.$$

This measure of change, $\Delta_{t-1,t}$, can be a difference (ratio) estimated using data only from the overlapping rotation groups, which is then added to (multiplied by) the composite estimate for period $t-1$. The composite estimate then becomes a recursively defined function of data collected in prior time periods:

$$Y_t^{CE} = (1-k)Y_t^{HT1} + kY_t^{HT2},$$

where $0 < k < 1$.

Composite estimators can often be expressed as a linear combination of simple estimates—one formed from each rotation group at each period. A few constraints are usually imposed. First, when estimating levels of a variable at time t, one usually requires that (a) the weighting coefficients of the group estimates at time t add to 1, and (b) for each period before t, the coefficients sum to 0. These restrictions ensure that no bias is introduced through the compositing. Second, to maintain the consistency of estimates, it is customary, at least for statistical agencies, to require that (a) the estimate of changes in a variable equal the difference (or ratio, for multiplicative composite estimators) of the appropriate estimates of levels for that variable, and (b) the estimates of components sum to the estimate of the corresponding total.

Composite estimation tries to take advantage of correlations over time. For example, suppose $x_{t-1,g}$ and $x_{t,g}$ are estimates from the same rotation group, g, for periods $t-1$ and t. If, due to sampling variability, $x_{t-1,g}$ is below its expected value, then $x_{t,g}$ tends to be as well. By assigning coefficients with opposite signs to the two estimates, one can temper the sampling variations while still balancing coefficients to ensure an unbiased estimate overall.

Variances and biases for composite estimators are computed according to the rotating panel design and depend on the variances and correlations of the rotation group estimates, which are often assumed to be nearly stationary over time. Thus, determining an optimal design becomes a problem of choosing the estimator's coefficients to minimize the expected error function. However, the problem becomes more complex when one considers the effect of the design on the different variables of interest, and on the several types of estimates to be disseminated: levels at specific points in time, changes across time, or averages over time. Changing the design or the estimators' coefficients to lower the expected error for a composite estimator of the level for a variable may induce a corresponding increase in the estimator for the change in a variable, and vice versa.

When the survey's most important estimate is a measure of the change in a variable over consecutive periods, a complete sample overlap is often the most efficient, as it makes the greatest use of the correlations over time. With a complete overlap, composite estimation with information from prior periods is generally not a consideration. However, for estimating the level at each time period, a partial sample overlap is often the most productive. Due to the constraint of consistency (see earlier discussion), when estimates of level and changes are both required, a compromise design may be used whereby a large fraction of the sample, but not all of the sample, is carried over from one period to the next.

Specific Examples of Composite Estimators

A specific example of a composite estimator is the one used in the Current Population Survey, jointly sponsored by the U.S. Bureau of Labor Statistics and the Census Bureau, to measure the U.S. labor force. In each month, separate estimates of characteristic totals are obtained from the eight rotation groups. Six of these groups contain households that were interviewed the prior month. The composite estimator implemented in 1998 combines the estimates from current and prior months to estimate the number of unemployed using one set of compositing coefficients, and the number of employed using a different set that reflects the higher correlations over time among estimates of employed:

$$Y_t^{CE} = (1-K)Y_t^{AVG} + K(Y_{t-1}^{CE} + \Delta_{t-1,t}) + Ab_t,$$

where Y_t^{AVG} is the average of the estimates of total from the eight rotation groups; $\Delta_{t-1,t}$ is an estimate of change based only on the six rotation groups canvassed at both times $t-1$ and t; b_t is an adjustment term inserted to reduce the variance of Y_t^{CE} and the bias arising from panel conditioning; and $(K, A) = (0.4, 0.3)$ when estimating unemployed, and (0.7, 0.4) when estimating employed. For researchers, a problem with composite estimates is producing them from public use microdata files, because computing the composite estimate for any period generally requires one to composite recursively over a number of past periods. This problem has been addressed for the Current Population Survey, which now produces and releases a set of "composite weights" with each month's public use file. First, for any month, composite estimates are determined for the labor force categories broken down into a number of race and ethnicity subgroups. Then, using these composite estimates as controls, the survey weights are raked to guarantee that the corresponding weighted estimates agree with the composite controls. The resulting composite weights can then be used to produce composite estimates simply by summing over the weights of records with the appropriate characteristics.

In the U.S. monthly surveys of retail and wholesale trade conducted before 1998 by the U.S. Census Bureau, a different rotating panel design led to an interesting set of composite estimators. In each of three consecutive months, one of three rotation groups was canvassed. In month $t+1$, businesses in rotation group A provided sales data for the months t and $t-1$, yielding estimates x_t^A and x_{t-1}^A, respectively. A preliminary composite estimate for month t,

$$P_t = (1-\beta)x_t^A + \beta P_{t-1}\Delta_{t-1,t},$$

was released, where $\Delta_{t-1,t} = x_t^A / x_{t-1}^A$, and $\beta = 0.75$ for the retail survey and 0.65 for the wholesale survey. One month later, firms in rotation group B supplied data for months $t+1$ and t, providing estimates x_{t+1}^B and x_t^B, respectively. This led to a final composite estimate for month t,

$$F_t = (1-\alpha)x_t^B + \alpha P_t,$$

where $\alpha = 0.80$ for the retail survey and 0.70 for the wholesale survey and an analogous preliminary estimate for month $t+1$. The third group was similarly canvassed a month later, and then the sequence was repeated. The difference between the final and preliminary composite estimates for month t, $F_t - P_t$, was called the revision in the estimate. In 1997 this rotating panel design was replaced by a complete sample overlap, due to problems of panel imbalance and differential response bias (early reporting bias) that led to undesirably large revisions in some months.

Different forms of composite estimators can be used to combine information from a survey and outside sources. In Statistics Canada's Labour Force Survey, the households in all six rotation groups are interviewed each month, with a new group entering and an old one dropping out each month. In any month, an estimate of total is obtained from each of the six groups. A composite regression estimator uses information from the six group estimates, Y_t^{AVG}; current population controls, X_t^{POP}; and composite regression estimates of the labor force from the prior month, Z_{t-1}^{CR}:

$$Y_t^{CR} = Y_t^{AVG} + [(X_t^{POP}, Z_{t-1}^{CR}) - (X_t^{AVG}, Z_t^{AVG})]b_t^{CR},$$

where the superscript AVG denotes an estimate based on data from the current survey period, and b_t^{CR} is the estimated composite regression parameter for month t. The estimation procedure guarantees accordance with the population controls, while taking advantage of recent labor force data. Using a different approach, Statistics Netherlands combines responses from demographic surveys and administrative data from social registers through regression estimation and a method called "repeated weighting" in order to reduce the variances of the estimators and to maintain numerically consistent tables across all official publications.

Patrick J. Cantwell

See also Current Population Survey (CPS); Panel; Panel Conditioning; Raking; Response Bias; Rotating Panel Design; Variance Estimation

Further Readings

Gambino, J., Kennedy, B., & Singh, M. P. (2001). Regression composite estimation for the Canadian Labour Force Survey: Evaluation and implementation. *Survey Methodology*, *27*(1), 65–74.

Houbiers, M. (2004). Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics*, *20*(1), 55–75.

Lent, J., Miller, S., Cantwell, P., & Duff, M. (1999). Effects of composite weights on some estimates from the Current Population Survey. *Journal of Official Statistics*, *15*(3), 431–448.

U.S. Census Bureau. (2006, October). *Technical paper 66: Current population survey, design and methodology.* Retrieved January 25, 2006, from http://www.census.gov/prod/2006pubs/tp-66.pdf

Wolter, K. M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, *74*, 604–613.

Comprehension

Survey researchers, in developing questions, must bear in mind the respondent's ability to correctly grasp the question and any response categories associated with the question. Comprehension, which is defined in this context as a respondent's ability to accurately understand a question and associated response categories, is crucial to reliable measurement of attitudes and behaviors.

Scholars have identified a number of elements in question wording that can interfere with comprehension: ambiguous language, vague wording, complex sentence structures, and presuppositions about the experiences of the respondent. The consequences of comprehension problems can be severe. If respondents' understanding of the question varies significantly from one respondent to another, the responses could provide a highly distorted picture of an attitude or behavior at the aggregate level.

Researchers have identified a number of techniques and guidelines to reduce the potential effects of question wording on comprehension:

1. Use clear, simple language in questions.
2. Use simple question structures, minimizing the number of clauses in a question.
3. Include a screening question if the survey is measuring attitudes or behaviors that might be unique to a specific group, and thereby skip all other respondents past the measures targeted to that group.
4. Provide definitions or examples in questions that may have terms that are ambiguous or vague.
5. Offer a frame of reference for terms that define a period of time (e.g., "in the past 7 days" as opposed to "recently").
6. Train interviewers to recognize problems with comprehension, and provide the interviewers with a uniform set of definitions and probes to address the problems.
7. Pretest survey questions not only with survey interviews, but in qualitative settings such as focus groups or in-depth cognitive interviews if resources permit.

Timothy Vercellotti

See also Cognitive Aspects of Survey Methodology (CASM); Cognitive Interviewing; Focus Groups; Pilot Test; Questionnaire Design; Reliability; Response Alternatives

Further Readings

Fowler, F. J., Jr. (1995). *Improving survey questions.* Thousand Oaks, CA: Sage.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context.* New York: Academic Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling, and data analysis* (3rd ed.). Thousand Oaks, CA: Sage.

Computer-Assisted Personal Interviewing (CAPI)

Computer-assisted personal interviewing (CAPI) refers to survey data collection by an in-person interviewer (i.e., face-to-face interviewing) who uses a computer to administer the questionnaire to the respondent and captures the answers onto the computer. This interviewing technique is a relatively new development in survey research that was made possible by the personal computer revolution of the 1980s.

Background

To understand the evolution of CAPI it is necessary to understand the history that led to its development and widespread implementation. In the late 1980s, many surveys used early versions of computer-assisted telephone interviewing (CATI). The early CATI systems ran as terminal applications on a mainframe or minicomputer. Computer applications typically used compilers; the central computer had to handle many simultaneous processes to service a CATI research facility. The cost of mainframes and more capable minicomputer systems was so high that the economic case that CATI should replace paper-and-pencil interviewing (PAPI) was tenuous. In addition, CATI facilities tended to use interviewers quite intensively and with close supervision, so interviewers tended to make fewer errors of the sort that computerized systems suppress, at least relative to face-to-face interviewers. With computing costs high, CATI was not a strong value proposition.

As personal computers (PCs) started to penetrate the market, they offered only modest processing power—but CATI interviews did not require much power. An intensively used PC could be cost-effective, and its capabilities matched the CATI task better than a mainframe or minicomputer did. There was no strong need to have a networked solution for PC computing, since CATI facilities could use low-tech case management and scheduling systems and still get the work done.

The PC software solutions for computer-assisted interviewing were adaptations of software first used on minicomputers or mainframes. A boundary constraint was that the compiler needed to have a variant that ran on DOS—the disk operating system for PCs that soon outstripped the use of Apple computers' proprietary operating system. This limited the software options.

By the late 1980s all major survey organizations doing face-to-face interviewing looked to establish a CAPI capability. With limited computing power for laptop computers and the limitations of DOS (which limited executable size because of its address space), these organizations faced a daunting systems challenge. Designers had two major strategic software alternatives. One choice was to follow the existing strand of software development with CATI and program the instrument to run on a laptop, accepting the reductions in memory and processing speed imposed by the technology of the times. The second strategic strand was to represent the instrument not as program code to execute but as a series of data records to be processed one by one. Internal machine instructions became records to be processed in exactly the same way, except that there was no output to the screen. The first application of this second strategy was done by Willem Saris of the Netherlands for smaller, less complex market research and public opinion surveys. In 1989, the Center for Human Resource Research at Ohio State University used a CAPI system based on representing the instrument as data to administer Round 11 of the National Longitudinal Survey of Youth 1979 (NLSY79), a large, complex event history interview that collected socioeconomic data in a one-hour face-to-face interview.

Weakness and Benefits

While case management is important in face-to-face interviews, there is no compelling reason other than marketing strategy by vendors to integrate the data capture engine for CAPI with case management. The two processes are logically separable. Indeed, in the early days of CAPI, the case management systems were rudimentary, and the survey process went ahead with no problems, as it had for decades before.

The weakness of the current standard CAPI strategy is that it is based on a computing paradigm that is 2 decades old. The current standard for computing emphasizes two things: (1) the use of modern relational databases, and (2) the use of the Web, especially coupled with relational database technology. CAPI systems based on relational databases and Web technology have several advantages. First, they integrate with parts of the survey process for which integration is compelling. Second, they can exploit systems tools that service a variety of data processing applications instead of requiring survey organizations to write de novo auxiliary utilities for their CAPI systems. Third, they provide a simple path toward implementing multi-modal and multi-platform surveys. Fourth, question records can be reused and reshuffled, thus speeding the design and modification of an instrument.

CAPI changes the survey process in many ways, but perhaps the most important way is that it forces a great deal of preparation to come early in the process. With PAPI, one only had to type up a printed questionnaire that interviewers could follow. While the data were being collected, the central office could put together a plan for processing and preparing the data. With CAPI, one must specify every action to be taken under every interview contingency. This fully contingent interview form must guide the interviewer through every step of the interview, and it must be ready in time for complete testing and the preparation of training materials. This front-loads the work process to such a degree that once the survey is in the field, most of the processing work is done.

The programming versus database paradigm has implications for survey preparation. When the database approach is used, the preparatory work can be handled by a survey specialist rather than a programmer. With the instrument driven by data tables, the authoring process is primarily a matter of filling in the blanks on a form. With the programming approach, the survey specialist has to communicate with the programming staff, increasing the chances for confusion, error, and miscommunication.

Usage

When it comes to the field effort, it is important to remember that, more and more, survey efforts are multi-modal. Face-to-face surveys frequently work many of their cases over the phone, self-administered on the Web, by mail, or even self-administered on a personal digital assistant (PDA) or some other device. Unless the technical approach handles multi-modal surveys efficiently, the survey preparation phase will require a separate programming effort for each mode. Apart from the multi-modal aspects, whatever system is on the computer must be used by interviewers, many of whom do not have a technical background. Sometimes programmers forget this. The key to a successful CAPI effort is simplicity. For example, when interviewers were trained in 1989 when the NLS became the first longitudinal survey to conduct a CAPI interview, the keystone of the training sessions was three words: "Read the screen." By breaking a complex interview into a few simple question types that one used over and over, it was relatively easy to train the interviewers.

Nearly 20 years later, the Web has penetrated the market with near ubiquity. By adopting a standard Web interface for CAPI systems, chances are improved that the interviewers who are recruited will be familiar with the look and feel of the application. As wireless connections over the cellular network spread and become more capable, survey research organizations have begun to interview with laptops connected to the central office over the cellular network. This integrates the field effort around the central office, bringing the field full circle to where CAPI began with a central facility serving the interviewer who is working on what is, essentially, a terminal.

Once the interviewer completes a case, the system must transmit the files to the central office. With the programming approach, one must generate specifications for this process. Done incorrectly, some data simply come up missing. With the database approach, each question record processed generates an answer record, and that answer record gets loaded into the master database used to design the survey, integrating the data and the documentation in a single resource. Regardless of the method, this integration needs to be achieved before researchers can use the data.

Surveys are all about creating databases, and for all but the most simply structured surveys (every respondent gets asked every question), the data set will have a variety of relationships that hold among the survey responses. Researchers collect data to analyze, and having a system built around a relational database to represent all parts of the questionnaire makes it easy to move the data into SAS, Statistical Package for the Social Sciences (SPSS), STATA, or some other statistical package. In the 1989 fielding of the NLSY79, the Ohio State system automatically produced SAS and SPSS control statements that read the data from the field—a capability that was years ahead of other systems. In recent years, much has been made of the Data Documentation Initiative (DDI) to provide a systematic method of documenting survey data sets that is reasonably similar across surveys. This would be done via Extensible Markup Language (XML)-formatted data for the survey questions. Ironically, the database approach to CAPI enabled this approach to documentation more than 15 years ago and, because the relational database tables needed to execute a survey are so comprehensive, even the questionnaire tables will contain documentation attributes at the question level that are far superior to DDI. With a database-designed system, one can load the data from a survey into a DDI-like system with minimal effort.

When it comes to disseminating the data, having the data already loaded into a relational database makes it relatively easy to produce a Web interface that allows users to search the database, peruse the codebook, and extract the desired data. Other techniques make this a case-by-case implementation of the necessary steps. Increasingly, major surveys are storing their data in relational databases for storage and manipulation, so the question becomes whether to take that step from the beginning or at the end of the process. Wireless methods will re-center CAPI around the Web and high-speed and highly secure central servers, greatly simplifying the technical support of field interviewers.

Randall Olsen and Carol Sheets

See also Computer-Assisted Telephone Interviewing (CATI); Face-to-Face Interviewing; Multi-Mode Surveys; Paper-and-Pencil Interviewing (PAPI)

Further Readings

Costigan, P., & Thomson, K. (1992). Issues in the design of CAPI questionnaires for complex surveys. In A. Westlake et al. (Eds.), *Survey and statistical computing* (pp. 147–156). London: North Holland.

Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nichols, W. L., et al. (Eds.). (1998). *Computer assisted survey information collection.* New York: Wiley.

Forster, E., & McCleery, A. (1999). Computer assisted personal interviewing: A method of capturing sensitive information. *IASSIST Quarterly*, *23*(2), 26–38.

Olsen, R. J. (2004). Computer assisted personal interviewing. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (Vol. 1, pp. 159–161). Thousand Oaks, CA: Sage.

Saris, W. E. (1991). *Computer-assisted interviewing*. Newbury Park, CA: Sage.

Computer-Assisted Self-Interviewing (CASI)

Computer assisted self-interviewing (CASI) is a technique for survey data collection in which the respondent uses a computer to complete the survey questionnaire without an interviewer administering it to the respondent. This assumes the respondent can read well (enough) or that the respondent can hear the questions well in cases in which the questions are prerecorded and the audio is played back for the respondent one question at a time (audio computer assisted self-interviewing—ACASI).

A primary rationale for CASI is that some questions are so sensitive that if researchers hope to obtain an accurate answer, respondents must use a highly confidential method of responding. For a successful CASI effort, the survey effort must consider three factors: (1) the design of the questions, (2) the limitations of the respondent, and (3) the appropriate computing platform.

Unless one has a remarkable set of respondents, the sort of instrument needed for CASI (or any self-administered interview) will be different from what one uses when a trained interviewer is administering the questionnaire. Having a seasoned interviewer handling the questioning offers a margin of error when designing questions. When the researcher is insufficiently clear, she or he essentially counts on the interviewers to save the day. Their help comes in a variety of forms. The interviewer can explain a question the respondent asks about. Good interviewing technique requires the interviewer to avoid leading the respondent or suggesting what the expected or "correct" response is. The interviewer can also help salvage bad questions when the respondent's answer reveals that, although the respondent showed no overt confusion about the question, it is clear that the respondent either did not understand the question or took the question to be something other than what was asked. The interviewer can also help out the questionnaire designer when, during a complex interview, it becomes clear to the interviewer that something has gone wrong with the programming and the item occurs in a branch of the questionnaire where it should not be. The interviewer can then try to put things right or at least supply a comment that will help the central office sort out the problem.

In all of these cases, the interviewer plays a crucial role in improving data quality. With a self-administered survey, regardless of mode, the safety net of a trained interviewer is not available. The circumstances of a self-administered interview put a real premium on clarity, computer assisted or not. The need for clarity is all the higher because there are no learning-curve effects for CASI—interviewers may do hundreds of cases, but with CASI essentially each respondent does just one. Thus, the question wording itself needs to be clear and self-contained so the respondent does not need to ask clarifying questions. Many surveys provide "help" screens to interviewers that have supporting information about a question, but that is not a good idea with CASI—using help screens violates the "Keep it simple" rule. Anything the respondent needs to see or read should be on the display screen the respondent sees, with no additional scrolling, clicking, or pressing of function keys. One should also be wary of question fills or elaborate skip patterns, since a simple error by the respondent can produce myriad problems that ripple through the rest of the instrument.

Because each respondent will see a question only once, designers must pay special attention to the layout of the screen. The first step the designer can take to reduce respondent confusion is to make sure similar questions have the same appearance. When the respondent is to pick the single best answer from a list, the screen should work in exactly the same way for each such question. If the respondent is to enter a date, he or she should either have to pick a day from a calendar or fill the date into the same sequence of data entry boxes, or if month and year is desired, use the same month/year format each time. If one introduces too many question styles and types, the opportunities for confusion escalate. The better choice is to rely on the minimum number of question types presented and structure the questionnaire so the respondent only has to deal with a very few question types utilized over and over.

With self-administered questionnaires, respondent satisficing behavior may arise. One can keep the attention of the respondent by using appropriate graphics if they help illustrate the concept. In some cases the problem is not keeping the respondent's attention but dealing with a limited attention span or even a limited ability to read. To handle these problems, audio computer-assisted self-interviewing helps both engage the respondent and minimize literacy problems. Research Triangle Institute (RTI) is the acknowledged pioneer in this area, having deployed ACASI in 1995 for the National Survey of Family Growth with feasibility tests prior to that. Today, ACASI is common with audio text fills and alternate language versions to adapt to respondents whose first language is not English. As equipment becomes smaller and more capable, survey researchers are beginning to set up ACASI interviews on Palm Pilots or other handheld devices. Ohio State University recently deployed an ACASI interview on Palm Pilots using an interview that took about an hour to complete—an interview full of very sensitive questions. The process went very smoothly; the respondent wore headphones to hear the question and tapped on the answer with a stylus on the Palm's screen. Respondents whose reading skills are strong can choose to turn off the audio. Because no one can tell whether the respondent is using audio, there is no stigma to continuing to use it.

Most interviews, however, do not consist only of sensitive questions. By putting the sensitive questions in one section, one can often switch between modes, using a CASI or ACASI method only where it is necessary. In fact, interviewers have been utilizing CASI since they first walked into a household with a computer, just as interviewers have turned a paper-and-pencil interviewing (PAPI) document into a self-administered questionnaire when they thought circumstances required it. For example, when a questionnaire asked about former spouses and the current spouse was in the house, savvy interviewers would simply point to the question in the booklet and say something like "And how would you answer this one?" With a laptop, the interviewer would simply twist the machine around and have the respondent enter an answer.

There are differences when using CASI within a telephone interview. One can conceal the line of sensitive questioning from another person in the room by structuring the questionnaire to require simple "Yes" or "No" responses, simple numbers, and the like. While this affords some confidentiality from eavesdropping, it does nothing to conceal the respondent's answers from the interviewer. To work on this problem there has been some limited experimentation at Ohio State using Voice over Internet Protocol (VoIP) methods. With some sophisticated methods, one can transfer the respondent to a system that speaks the questions by stringing together voice recordings and then interprets the respondent's answers and branches to the appropriate question. When done with the sensitive questions, the "robot" reconnects the interviewer and the interview continues. This approach works quite well and allows telephone interviews to achieve a measure of the security attained with other "closed" methods, although there has yet to be a controlled experiment that has compared VoIP effectiveness with results achieved by traditional CASI or ACASI techniques.

Randall Olsen and Carol Sheets

See also Audio Computer-Assisted Self-Interviewing (ACASI); Computerized Self-Administered Questionnaires (CSAQ); Paper-and-Pencil Interviewing (PAPI); Satisficing; Voice Over Internet Protocol (VoIP) and the Virtual Computer-Assisted Telephone Interview (CATI) Facility

Further Readings

Turner, C. F., Villarroel, M. A., Chromy, J. R., Eggleston, E., & Rogers, S. M. (2005). Same-gender sex in the USA: Trends across the 20th century and during the 1990s. *Public Opinion Quarterly*, *69*, 439–462.

Turner, C. F., Villarroel, M. A., Rogers, S. M., Eggleston, E., Ganapathi, L., Roman, A. M., et al. (2005). Reducing bias in telephone survey estimates of the prevalence of drug use: A randomized trial of telephone Audio-CASI. *Addiction*, *100*, 1432–1444.

Villarroel, M. A., Turner, C. F., Eggleston, E. E., Al-Tayyib, A., Rogers, S. M., Roman, A. M., et al. (2006). Same-gender sex in the USA: Impact of T-ACASI on prevalence estimates. *Public Opinion Quarterly*, *70*, 166–196.

Computer-Assisted Telephone Interviewing (CATI)

Computer-assisted telephone interviewing (CATI) in its simplest form has a computer replacing the paper questionnaire on a telephone interviewer's desk.

Advantages of Computer-Assisted Telephone Interviewing

CATI provides the following advantages:

- More efficient data collection, because the interviewer enters answers directly into the computer rather than sending a paper questionnaire for a separate data capture step.
- More efficient and more accurate questionnaire administration, because the computer delivers the questions to the interviewer in the correct programmed sequence, including any required rotations, randomizations, or insertions of information from a separate data file or from earlier in the interview.

- More accurate data collection, because the computer can apply various range and logic edits as the answers are entered. These edits can range from hard edits (in which the system will not accept an answer outside certain parameters—for example, age at first marriage being less than 14 years of age) to "query edits" that require the interviewer to confirm that, while unusual, the answer is indeed that intended by the respondent (e.g., to confirm that age at first marriage was indeed only 14 years of age).

While this has been the basic model for CATI systems since they were first introduced in the 1970s, and some CATI systems still have only this questionnaire administration component, technological developments during the past 30 years have provided many more ways in which the computer can assist the telephone interviewing process.

Quality Assurance Monitoring

For quality assurance, most telephone surveys have a sample of interviews monitored by a supervisor, so the researcher can be confident that the questions have been administered by the interviewer as instructed (correct wording, probing) and the answers given by the respondent faithfully recorded or correctly categorized. Computers allow this to be done in an unobtrusive and effective manner, usually by the supervisor listening in on the interview on a separate audio channel while watching an image of the interviewer's screen.

Further assistance by the computer for this process occurs with the automatic recording of the interviewer the supervisor is monitoring and for what time period. A data entry tool for the supervisor then records the results of the monitoring session and a database in which these results are stored. The use of better allocation of monitoring resources, typically by an algorithm, queries the database, so that more experienced interviewers who rarely have errors are monitored less than those who are newer or who have been identified as needing more assistance.

Sample Management and Call Scheduling

Most CATI programs now have at least two modules, one being the questionnaire administration tool already described, the other providing sample management and call scheduling functions, such as the following:

- Holding the list of all the telephone numbers to be called, along with any other relevant frame information, for example, geographic region if the sample is to be stratified by region
- Recording information about the call history, that is, each call made to each number, such as time and date the call was placed, the interviewer who placed the call, and the call outcome (completed interview, refusal, busy signal, etc.)
- Executing calling rules that determine when the next call (if any) should be placed to a number, which could include delays from the previous call, or certain times of day or parts of week
- Prioritizing among numbers competing for delivery at the same time, for example, by queuing numbers that have appointments first, calls to households where previous contact has occurred next, and fresh sample last
- Delivering phone numbers to the next available interviewer appropriate for that number (e.g., previous refusals to refusal converter interviewers)
- Producing sample progress information, such as number of interviews so far completed by strata, number of interviews refused, and amount of sample yet to be worked

The sample management module often has a separate supervisor interface, which enables the supervisor to execute additional sample management functions, such as stopping particular numbers from being delivered to increasing for a limited period of time the priority of numbers in strata where the survey is lagging.

Automated Dialing and Other Call-Handling Assistance

Telephone technology, typically with a separate computer residing in the PBX (private branch exchange) or dialer, can also be considered part of a CATI system. While the main drivers of telephone technology have been telemarketing and other call centers, they still provide assistance to the telephone survey process by the following features:

- Autodialing, in which the actual act of dialing is performed on some trigger (such as a keystroke instruction from an interviewer, the interviewer logging in to the system or hanging up from the previous caller, or in the case of predictive dialers, when the probabilities of both an interviewer becoming free and a call resulting in a connect exceed some threshold)

- Auto-dispositioning, where the outcome of certain types of calls (e.g., busy, fax, disconnected) can be detected from the signal tones and coded by the dialer rather than by the interviewer
- Interactive Voice Response, or IVR, where a prerecorded voice replaces the interviewer and data is collected either by the respondent's key strokes or machine-recognizable words and phrases
- Automatic Call Distribution, or ACD, which organizes incoming calls into queues and delivers them to interviewers according to rules relating to call type and interviewer attribute
- Message push-out, in which the dialer can call numbers without any interviewer involvement and deliver pre-recorded messages to any person, voicemail, or answering machine that answers the call
- Recording of interviews for more accurate verbatim data capture or for more effective coaching of interviewers
- Playing of sound clips to the respondent (although these can also be stored in the questionnaire administration tool)

While some dialers have some sample management and basic questionnaire administration capabilities, at the time of writing there are few systems that manage the sophistication in questionnaire administration or sample management that is typically needed in survey work.

Network and Internet Issues

Most CATI systems use networked computers so that all interviewers working on the same survey share a single pool of telephone numbers, access the same version of the questionnaire, and all data is stored in a central database. There are many advantages of a network system over separate laptops or other personal computers. One advantage is centralized control over the survey instrument, so that mid-survey changes to the questionnaire can be instantaneously implemented to all terminals. Centralized control of the sample and data is also advantageous in that the risks of exceeding targets or not identifying problem areas quickly enough are minimized, and ensuring appropriate data backups are made. Network systems also facilitate supervision and monitoring functions.

The Internet provides additional assistance by allowing the use of Voice Over Internet Protocol to carry the audio channel rather than needing multiple phones connected into a limited number of PBX exchanges. This simplifies wiring needs in centralized CATI centers and enables distributed virtual call centers, through which interviewers can work from their homes as long as they have a sufficiently fast Internet connection.

The Future of Computer-Assisted Telephone Interviewing

Benefits

When compared with the three other main modes of survey data collection (Web, personal interviewing, mail), CATI still retains two advantages. First, it enables interviewer administration of questionnaires rather than self-completion, as required by Web and mail surveys. While there are situations in which self-completion can be methodologically preferable (for example, when collecting data on very sensitive topics), interviewer-administered surveys typically carry the advantages of higher response rates, higher item completion rates, and the opportunity to probe the respondent to get more complete answers. The second advantage is that when compared with the other interviewer-administered mode—face-to-face interviewing—CATI is typically more cost-effective and provides for faster delivery of data.

Challenges

There are, however, challenges to CATI surveys that require resolution if CATI surveys are to retain more advantages relative to their disadvantages. One such challenge is the proliferation of cell phones (in many cases replacing landlines completely in households) combined with societal and sometimes legal restrictions on the extent to which cell phones can be used in surveys. Legislative restrictions also influence telephone surveys; some states include telephone surveys in the scope of "do-not-call" restrictions, and others restrict the use of some features on the more advanced automated dialers. Although such legislation is aimed more at reducing invasion of privacy by telemarketers, and in some cases specifically excludes legitimate survey research from the restrictions, the distinction between telemarketing and survey research often is not recognized at the household level. Another challenge is the increasing reluctance of the public to participate in telephone surveys, although the presence of "do-not-call" lists and other privacy-protecting measures may in fact work to the advantage of CATI surveys to the extent they will reduce telemarketing and other nuisance calls that have led to the current resentment of telephone intrusion in households. The apparently significantly lower cost of Internet surveys compared with CATI surveys also creates a challenge, although, as noted earlier, there are methodological issues that still work in favor of CATI surveys.

Jenny Kelly

See also Do-Not-Call (DNC) Registries; Interviewer Monitoring; Outbound Calling; Paper-and-Pencil Interviewing (PAPI); Predictive Dialing; Sample Management; Voice Over Internet Protocol (VoIP) and the Virtual Computer-Assisted Telephone Interview (CATI) Facility

Further Readings

Hansen, S. E. (2008). CATI sample management systems. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 340–358). New York: Wiley.

Kelly, J., Link, M., Petty, J., Hobson, K., & Cagney, P. (2008). Establishing a new survey research call center. In J. Lepkowski et al. (Eds.), *Advances in telephone survey methodology* (pp. 317–339). New York: Wiley.

Steve, K., Burks, A. T., Lavrakas, P. J., Brown, K., & Hoover, B. (2008). The development of a comprehensive behavioral-based system to monitor telephone interviewer performance. In J. Lepkowski et al. (Eds.), *Advances in telephone survey methodology* (pp. 401–422). New York: Wiley.

Computerized-Response Audience Polling (CRAP)

A number of survey designs deviate from the parameters of a scientific probability design, with significant consequences for how the results can be characterized. Computerized-response audience polling (CRAP) is an example of such a design. In this kind of poll, a sample of telephone numbers is typically purchased and loaded into a computer for automatic dialing. The questionnaire is produced through computer software that employs the digitized voice of someone assumed to be known to many of those who are sampled, such as the voice of a newscaster from a client television station. After an introduction, the computerized voice goes through the questionnaire one item at a time, and the respondent uses the key pad on a touchtone phone to enter responses to each question asked, as in an interactive voice response (IVR) system.

A major problem with CRAP polls is that the methodology does not allow for specific respondent selection, meaning that the basic premise of probability sampling, namely that each respondent has a known, nonzero probability of selection, is violated. Interviews are conducted with whoever answers the phone, and there is no guarantee that the person answering is eligible by age or other personal characteristics. Although information can be gathered about the household composition, there is no random selection of a designated respondent from the household. The computer can dial a large set of telephone numbers in a short period of time, working through a purchased sample quickly but producing a low contact or cooperation rate as a result. There also is no attempt to recontact a household to obtain an interview with a designated respondent who is not at home at the time of the first call. Because of these considerations, it is inappropriate to calculate a margin of error around any estimates produced from such a poll.

This method shares many characteristics with self-selected listener opinion polls (SLOP) and other designs that employ volunteer samples. A true response rate cannot be calculated, although a version of a cooperation rate can. The data can be collected rapidly and at low cost. Although post-stratification weighting can be applied to the resulting set of respondents, it is difficult to interpret its meaning when information about respondent selection is missing.

Michael Traugott

See also Interactive Voice Response (IVR); Mode of Data Collection; Self-Selected Listener Opinion Poll (SLOP)

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Computerized Self-Administered Questionnaires (CSAQ)

Computerized self-administered questionnaires (CSAQ) are a method of collecting survey data that takes advantage of computer technology to create an instrument (the questionnaire) that allows respondents to complete the survey with little or no other human assistance. Applications range from completely self-administered questionnaires to the use of data collectors who provide introductory information and technical assistance if needed. CSAQ applications include Web surveys in which respondents go to a designated Web site and complete the survey online; research in public access areas in which a respondent may answer questions presented at a kiosk or on a computer provided by a vendor at a conference or convention; touchtone data entry such as telephone surveys in which the respondents use the telephone keypad to enter their responses; and surveys in which the use of CSAQ is one portion of the overall interview process. Surveys of this type are also called "computer-assisted self-administered personal interviewing," "computer-assisted self-administered interviewing (CASI)," or "audio computer-assisted interviewing (ACASI)."

The use of CSAQ has several advantages over traditional self-administered paper-and-pencil (PAPI) surveys. It allows the use of complex skip patterns, directing respondents to the next appropriate question based on an answer or answers to previous questions. It also allows questions to be "personalized" based on demographic variables such as age, race, or sex; or use answers provided earlier in the questionnaire as part of wording of questions coming later. For example, knowing the sex of a child would allow a subsequent question wording to ask about the respondent's "son" or "daughter" rather than his or her "child." The use of CSAQ can be helpful in surveys that ask sensitive questions about sexual activity or drug use for which respondents might be hesitant to provide such information to an interviewer either face to face or over the telephone.

CSAQ designs that use devices with a video monitor and speakers (such as a laptop, monitor, or kiosk) can include graphics such as pictures and illustrations. Using speakers or headphones, audio clips can also be added. Video clips can be used to illustrate a product or to screen an advertisement or public service announcement. Audio clips can be used to "read" the questionnaire in those designs in which the target population may be illiterate or have limited reading ability.

Other advantages of CSAQ designs include reducing the cost of a survey (because interviewers may not be needed) and minimizing data entry errors (because the responses are entered directly into a database at the time the survey is completed). This can reduce the amount of time needed to verify the data and complete the analysis.

The major disadvantages of using CSAQ involve the design of the survey instrument. It must be designed in such a way that the questionnaire flows smoothly. Respondents, especially those who are less comfortable with the use of computers, may become easily frustrated with a questionnaire that is not self-explanatory or in which the questions are not easily understood. The visual layout will influence not only the response rate but the quality of data as well. Special attention must be paid to issues such as font size, color combinations, page layout, and the method used for the respondents to record their answers (radio button, number, open-ended). Web-based CSAQ must be designed in such a way that they are compatible with the variety of screen resolutions and Web browsers that are in use.

As with any survey, sample bias is a consideration. This is especially true for CSAQ designs that make no attempt to identify, screen, or select respondents on some random basis. While results from such a survey may be useful for some purposes, explicit reference must be made of the limitations of drawing any conclusions from the results.

Computerized self-administered questionnaires can be a powerful tool to improve the quality and reduce the cost of survey data collection. However, as with any survey research method, the researcher must consider the limitations of the method used and attempt to reduce or eliminate the effects of those limitations.

Dennis Lambries

See also Audio Computer-Assisted Self-Interviewing (ACASI); Computer-Assisted Self-Interviewing (CASI); Internet Surveys; Paper-and-Pencil Interviewing (PAPI); Self-Selection Bias; Sensitive Topics; Touchtone Data Entry; Web Survey

Further Readings

Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nicholls, W. L., II, et al. (Eds.). (1998). *Computer assisted survey information collection.* New York: Wiley

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: Wiley.

CONFIDENCE INTERVAL

A probability sample can provide a point estimate of an unknown population parameter and the standard error of that point estimate. This information can be used to construct a confidence interval to give an estimated range of values around the point estimate that is likely to include the unknown population parameter.

For example, assume that a soda can–filling plant fills soda cans at an average rate of 1,000 to 1,500 cans per minute. Several filling nozzles are simultaneously used to fill the cans. Electronic sensors are used to ensure that the filled amount is within specified limits. Due to inherent variability in the filling process, it is impossible to fill an exact amount (355 milliliters [ml]) of soda in each can. As a final quality control measure, a quality assurance inspector wants to estimate the mean amount of soda filled in one particular batch of 120,000 cans. To do so, one extreme option would be to open all the cans and measure the contents. Clearly, this approach is not cost-effective because doing so will destroy all the cans and contaminate the soda. A reasonable approach would be to take a random sample of, say, 20 cans, measure their contents, and calculate the average amount of soda in each can. In survey sampling terminology, this average is known as the "sample mean." The average amount of soda in each of the 120,000 cans is called the "population mean."

It is a common practice to use the sample mean as a point estimate of the population mean. Suppose the sample mean is calculated to be 352 ml. Does it make sense to infer that the population mean also is 352 ml? If "Yes," then what is the margin of error in drawing such an inference? If another random sample of 100 cans yields a sample mean of 355.8 ml, then the inspector will have more confidence in making an inference about the population mean as compared with an inference based on a random sample of 20 cans because she or he will be using more information in the inference. If the inspector had additional information that the filled amount of soda does not vary much from can to can (i.e., information that the population standard deviation of the filled amount of soda is quite small), then a random sample of 20 cans may be sufficient to draw a conclusion about the population mean with reasonable confidence. On the other hand, if the filled amount of soda varies a lot from can to can (i.e., the population standard deviation of the filled amount is very large), then even a random sample of 100 cans may not be sufficient to draw any conclusion about the population mean with desired confidence.

This example shows that point estimates alone are not sufficient for drawing conclusions about a population characteristic unless accompanied by some additional information regarding the level of confidence and margin of error involved in the estimation process. It would be more informative if the inspector could make a statement, such as "I am 95% confident that, on average, between 354.5 ml to 355.3 ml of soda is present in the 120,000 cans." Such statements are facilitated by adopting the method of confidence intervals for estimation or statistical inference purposes.

Detailed Definition of a Confidence Interval

In statistical terms, a confidence interval (two-sided) is defined as a random interval $[L, U]$ enclosing the unknown population parameter value (θ) (such as a population mean, variance, or proportion) with a given probability $(1-\alpha)$. That is, Probability $(L \leq \theta \leq U) = 1 - \alpha$, where $0 \leq \alpha \leq 1$ and it generally takes small values, such as 0.01, 0.05, or 0.1. The interval $[L, U]$ is known as the $100(1-\alpha)\%$ confidence interval for θ, and the probability $(1-\alpha)$ is known as the confidence level or the coverage probability of the interval $[L, U]$. In certain applications, only a lower or upper bound may be of interest, and such confidence intervals are known as "one-sided" confidence intervals.

If a sampling process is repeated a large number of times, and for each selected sample a confidence interval is obtained using the same confidence level and statistical technique, and if the population parameter was known, then approximately $100(1-\alpha)\%$ of the confidence intervals will enclose the population parameter. In reality, θ is unknown, and owing to budget and time constraints, only one sample is selected; hence, it is not possible to know with certainty if the calculated $100(1-\alpha)\%$ confidence interval encloses the true value of θ or not. It is hoped with the chances at $100(1-\alpha)\%$ that it does enclose the true value of θ. The lower and upper end points of the confidence interval depend upon the observed sample values, selected confidence level, statistical technique, the sample design, and population distributional characteristics, as illustrated by the following examples. The confidence interval definition given earlier comes from the *frequentist* school of thought. The alternative, Bayesian inference, is not yet commonly used in survey data analysis and hence is not covered here.

Construction of a Confidence Interval

Let $\hat{\theta}$ denote an estimator of θ and $v(\hat{\theta})$ denote its variance, then a $100(1-\alpha)\%$ confidence interval is given by $\hat{\theta} \pm c\sqrt{v(\hat{\theta})}$, where c is a constant such that

$$\text{Probability}\left(-c \leq \frac{(\hat{\theta}-\theta)}{\sqrt{v(\hat{\theta})}} \leq c\right) = 1-\alpha,$$

where the probability is calculated using the sampling distribution of $\hat{\theta}$. In most cases, the sampling distribution of $\hat{\theta}$ is not known and is assumed to be either a normal (Gaussian) or Student's t-distribution depending upon the sample size and distributional characteristics of the population. If $v(\hat{\theta})$ is also not known, then its estimated value, $\hat{v}(\hat{\theta})$, is used in the calculation. Due to these reasons, the confidence interval obtained will not be exact (i.e., the coverage probability will be close to $1-\alpha$).

In nonsurvey data analyses, confidence intervals are calculated based on the assumption that simple random sampling with replacement was used, or equivalently, that the random sample was selected from an infinite population. However, in most surveys the target population is finite, and a more complex sampling scheme is used to sample the finite population. Hence, the usual central limit theorem cannot be applied to the finite population sampling cases. Instead, a central limit theorem proposed by Jaroslav Hájek is used for approximating the sampling distribution of $\hat{\theta}$ by a normal distribution for sufficiently large sample sizes.

For the following examples, suppose U denotes the finite population consisting of N units $\{y_1, y_2, \ldots, y_N\}$ and Σ denotes a random sample of n units $\{y_1, y_2, \ldots, y_n\}$ selected from U. Let

$$Y = \left(\sum_{i=1}^{N} y_i\right), \bar{Y} = N^{-1}Y, \text{ and}$$
$$S^2 = (N-1)^{-1}\left(\sum_{i=1}^{N} y_i - \bar{Y}\right)^2$$

be the unknown population total, mean, and variance of the N units of the population U, respectively. Similarly, let

$$y = \left(\sum_{I=1}^{n} y_i\right), \bar{y} = n^{-1}y, \text{ and}$$
$$s^2 = (n-1)^{-1}\left(\sum_{i=1}^{n} y_i - \bar{y}\right)^2$$

be the sample total, mean, and variance of n units of sample S, respectively.

Simple Random Sampling Without Replacement

It is well known that the sample mean $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$ and the variance of $\bar{y}$ is $v(\bar{y}) = n^{-1}(1-\frac{n}{N})S^2$. Then, an approximate $100(1-\alpha)\%$ confidence interval for $\bar{Y}$ is given by $\hat{\theta} \pm c\sqrt{v(\hat{\theta})} = \bar{y} \pm c\sqrt{v(\bar{y})}$. If the sample size n is large enough to satisfy the assumptions of Hájek's central limit theorem, then the sampling distribution of $\frac{(\hat{\theta}-\theta)}{\sqrt{v(\hat{\theta})}} = \frac{(\bar{y}-\bar{Y})}{\sqrt{v(\bar{y})}}$ can be approximated by a standard normal probability distribution function with $mean = 0$ and $variance = 1$. The value of c can be obtained by solving the equation Probability $\left(-c \leq \frac{(\bar{y}-\bar{Y})}{\sqrt{v(\bar{y})}} \leq c\right) = 1-\alpha$. By applying elementary statistics results, $c = z_{\alpha/2}$, where $z_{\alpha/2}$ is the 100 $(1-\alpha/2)th$ percentile of the standard normal distribution. Hence, an approximate large sample $100(1-\alpha)\%$ confidence interval for $\bar{Y}$ is given by $\bar{y} \pm z_{\alpha/2}\sqrt{v(\bar{y})}$. Note that $v(\bar{y})$ involves the unknown population variance S^2, which is estimated by the sample variance s^2. If the sample size is not large enough to ensure asymptotic normality, and it is reasonable to assume that the population units follow a normal distribution, and if S^2 is unknown, then the sampling distribution of $\frac{(\bar{y}-\bar{Y})}{\sqrt{v(\bar{y})}}$ can be approximated by a Student's t-distribution with $(n-1)$ degrees of freedom. In that case, $c = t_{n-1,\alpha/2}$, where $t_{n-1,\alpha/2}$ is the $100(1-\alpha/2)th$ percentile of the Student's t-distribution function with $(n-1)$ degrees of freedom.

In the original example regarding soda cans, suppose a sample of 20 soda cans is selected using the simple random sampling without replacement (SRSWOR) methodology. Let y_i be the amount of soda in the I–th can, where $i = 1, 2, \ldots, 20$. Using the amount of soda in each of the 20 cans, $\bar{y}$ and s^2 can be calculated. Let us assume that $\bar{y} = 351$ ml and $s^2 = 25$ ml^2, then $v(\bar{y}) = n^{-1}(1-\frac{n}{N})S^2 \simeq n^{-1}s^2 = 25/20 = 1.25$ because the sampling fraction $n/N = 20/120{,}000$ is negligible in this case. In this example, it is reasonable to assume that the amount of soda in each of the 120,000 cans follows a normal probability distribution. Hence, an approximate 95% confidence interval for the mean amount of soda in the 120,000 cans is given by $351 \pm t_{n-1,\alpha/2}\sqrt{1.25}$. For a 95% confidence interval, $\alpha = .05$ and from the Student's t-probability distribution tables $t_{n-1,\alpha/2} = t_{19,.025} = 2.093$; hence, the 95% confidence interval is $351 \pm t_{n-1,\alpha/2}\sqrt{1.25} = 351 \pm 2.093\sqrt{1.25} = [348.66, 353.34]$.

Stratified Sampling

Suppose that $N = 120{,}000$ cans were produced in three batches of 40,000 cans each. In order to account for batch-to-batch variability in the estimation process, it would make sense to select a random sample from each of the three batches. This is known as "stratified sampling." To find a confidence interval for the average amount of soda in the 120,000 cans ($\bar{Y}$), suppose the inspector took a SRSWOR sample of 40 cans from each of the three batches. In stratified sampling notation (from William Gemmell Cochran), the three batches are known as "strata," with $N_1 = N_2 = N_3 = 40{,}000$ denoting the stratum sizes and $n_1 = n_2 = n_3 = 40$ denoting the stratum sample sizes. From Cochran, an unbiased estimator of $\bar{Y}$ is $\bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{n_h}\bar{y}_h$, where $\bar{y}_h$ denotes the sample mean for the $h - th$ stratum and L denotes the number of strata in the population. The variance of $\bar{y}_{st}, v(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) n_h^{-1} S_h^2$ involves unknown stratum variances S_h^2, which are estimated by the corresponding sample variances, $s_h^2 = (n_h - 1)^{-1}(\sum_{i=1}^{n_h} y_{hi} - \bar{y}_h)^2$, where y_{hi} denotes the value of the $i - th$ unit in the $h - th$ stratum.

In the preceding example, a sample of 40 cans from each stratum may be sufficiently large to assume a normal probability distribution function for $\bar{y}_{st}$. Hence, an approximate $100(1-\alpha)\%$ confidence interval for $\bar{Y}$ is given by $\bar{y}_{st} \pm z_{\alpha/2}\sqrt{\hat{v}(\bar{y}_{st})}$, where $\hat{v}(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) n_h^{-1} s_h^2$. If stratum sample sizes are not large enough, then a Student's t-distribution with n^* degrees of freedom is used to approximate the sampling distribution of $\bar{y}_{st}$. The calculation of n^* is not straightforward and should be done under the guidance of an experienced survey statistician.

Cluster Sampling

Now suppose that the $N = 120{,}000$ cans were packed in 12-can packs and shipped to a retailer. The retailer is interested in knowing the average amount of soda in the 120,000 cans ($\bar{Y}$). A stratified sampling would not be feasible unless all of the 12-can packs ($M = 10,000$) are opened. Similarly, for SRSWOR, it would require one to list all the 120,000 cans, which in turn may require opening all the packs. Because each pack can be regarded as a cluster of 12 cans, a cluster sample design is most suitable here. To obtain an approximate $100(1-\alpha)\%$ confidence interval for $\bar{Y}$, the retailer decided to select a SRSWOR sample of $m = 20$ packs from the population of $M = 10,000$ packs and measure the amount of soda in each of the cans. This is known as a single-stage (or one-stage) cluster sample, in which all the clusters (packs) have the same number (12 cans in each pack) of elements (soda cans). An unbiased estimator of $\bar{Y}$ is $\bar{y}_{cluster} = \frac{1}{N}\left(\frac{M}{m}\right)\sum_{i=1}^{m}\sum_{j=1}^{r} y_{ij}$, where r is the common number of elements in each cluster and y_{ij} denotes the value of the $j - th$ element in the $i - th$ selected cluster. Let $t_i = \sum_{j=1}^{r} y_{ij}$, $s_t^2 = (m-1)^{-1}(\sum_{i=1}^{m} t_i - \bar{t})^2$, and $\bar{t} = m^{-1}(\sum_{i=1}^{m} t_i)$, then the variance of $\bar{y}_{cluster}$ is estimated by $\hat{v}(\bar{y}_{cluster}) = v(\frac{M}{N}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{r} y_{ij}) = \left(\frac{M}{N}\right)^2 v(\frac{1}{m}\sum_{i=1}^{m} t_i) = \left(\frac{M}{N}\right)^2 \frac{1}{m}(1 - \frac{m}{M})s_t^2$. If the number of clusters in the sample is large, then an approximate $100(1-\alpha)\%$ confidence interval for $\bar{Y}$ is given by $\bar{y}_{cluster} \pm z_{\alpha/2}\sqrt{\hat{v}(\bar{y}_{cluster})}$.

In the preceding example, $r = 12$ (because all 12 cans in a pack are examined) and t_i represents the total amount of soda in the $i - th$ selected pack. Because a SRSWOR sample of $m = 20$ packs is not large enough to assume a normal probability distribution, a Student's $t-$distribution with $t_{m-1,\alpha/2} = t_{19,.025} = 2.093$ could be used by the retailer to obtain a 95% confidence interval for $\bar{Y}$ (i.e., $\bar{y}_{cluster} \pm t_{19,.025}\sqrt{\hat{v}(\bar{y}_{cluster})}$).

It is a common mistake to analyze the data obtained from a cluster sample as if it were obtained by SRSWOR. In the preceding cluster sampling example, if clustering is ignored at the analysis phase and a 95% confidence interval is constructed by assuming that the 240 (20 packs × 12) soda cans were selected using the SRSWOR method, then the actual coverage probability may be less than 95%, depending upon the size of the intracluster (or intraclass) correlation coefficient. Generally, the point estimate of a population parameter will be the same whether the data are analyzed with or without incorporating the survey design information in the estimation process. However, the standard errors may be quite different if the survey design information is ignored in the estimation process, which in turn will result in erroneous confidence intervals.

The examples given deal with constructing a confidence interval for the population mean for some of the basic survey designs. In practice, survey designs are generally more complex, and a confidence interval for other population parameters—such as population proportions and quantiles; linear, log-linear, and nonlinear regression model parameters; survival function at a given time (Cox's proportional hazard model or Kaplan-Meier estimator)—may be needed.

Akhil K. Vaish

See also Alpha, Significance Level of Test; Bias; Cluster Sample; Confidence Level; Finite Population; Inference; Margin of Error (MOE); Point Estimate; Population Parameter; ρ (Rho); Sampling Without Replacement; Simple Random Sample; Standard Error; Stratified Sampling; Target Population; Variance Estimation

Further Readings

Alf, C., & Lohr, S. (2007). Sampling assumptions in introductory statistics classes. *The American Statistician*, *61*(1), 71–77.

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Hájek, J. (1960). Limiting distribution in simple random sampling from a finite population. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, *5*, 361–374.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis.* Belmont, CA: Duxbury.

Research Triangle Institute. (2004). *SUDAAN language manual*. Release 9.0. Research Triangle Park, NC: Author.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Confidence Level

In statistical inference, it is common practice to report the point estimate of a population parameter along with its standard error (square root of the variance). Often, the point estimator and its standard error are combined by adding and subtracting from the point estimate a multiple of the standard error to obtain an interval estimator. Suppose $\hat{\theta}$ denotes an estimator of θ and $v(\hat{\theta})$ denotes its variance, then $\hat{\theta} \pm c\sqrt{v(\hat{\theta})}$ is an interval estimator of the parameter θ. The constant c is chosen in such a way that if the sampling process is repeated for a large number of times and for each sample an interval estimator is obtained, then approximately a pre-defined percentage of the intervals will enclose θ. This pre-defined percentage is known as the "confidence level" (or "coverage probability") of the interval estimator. Hence, interval estimators are also commonly known as "confidence intervals." In most cases, for a two-sided confidence interval, the value c is obtained by solving the equation

$$\text{Probability}\left(-c \le \frac{(\hat{\theta}-\theta)}{\sqrt{v(\hat{\theta})}} \le c\right) = 1-\alpha,$$

where $100(1-\alpha)\%$ is the chosen confidence level of the desired confidence interval and the probability is calculated using the sampling distribution of $\hat{\theta}$. Generally, the sampling distribution of $\hat{\theta}$ is not known and is assumed to be either a normal (Gaussian) or Student's t-distribution depending upon the sample size and distributional characteristics of the population. If $v(\hat{\theta})$ is also not known, then its estimated value, $\hat{v}(\hat{\theta})$, is used in the calculation.

If $\hat{\theta}$ is biased or $\hat{v}(\hat{\theta})$ is not calculated according to the sampling design or an incorrect sampling distribution of $\hat{\theta}$ is assumed, then the actual confidence level of the $100(1-\alpha)\%$ confidence interval will be different from the nominal confidence level $100(1-\alpha)\%$. For example, Carl-Erik Särndal, Benqt Swensson, and Jan Wretman examined the effect of bias on confidence level. Cherie Alf and Sharon Lohr showed that the true confidence level for a 95% confidence interval for the population mean may be less than 95% depending upon the intracluster correlation coefficient (i.e., if the sample design characteristics are ignored in the variance calculations, then the resulting confidence interval will not have the correct confidence level).

Akhil K. Vaish

See also Bias; Confidence Interval; Inference; Point Estimate; ρ (Rho); Variance Estimation

Further Readings

Alf, C., & Lohr, S. (2007). Sampling assumptions in introductory statistics classes. *The American Statistician*, *61*(1), 71–77.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

CONFIDENTIALITY

The confidentiality of survey data is expected by both survey researchers and survey participants. Survey researchers have multiple meanings for *confidentiality* that are not quite the same as the common definition. Dictionary definitions use terms such as *private, intimate*, and *trusted*, and some refer to national security concerns.

However, in survey research, the definition is more complex and can be used differently by different researchers and survey organizations. For the most part, confidentiality in survey research refers to the methods for protecting the data that are collected. It refers both to the promises made to survey participants that they will not be identified in any way to those outside the organization without their specific permission and to the techniques that organizations use to ensure that publicly available survey data do not contain information that might identify survey respondents.

For respondents, the promise of confidentiality is the agreement on the methods to prevent others from accessing any data that might identify them. Confidentiality of data is important for the success of survey research because survey participants would be much less willing to participate if they thought the survey organization would disclose who participated in the research and/or their identified responses to questions. The confidentiality protections provided to participants are not as strong as for anonymously collected data, but both anonymity and confidentiality are used for the same reasons.

The confidentiality of survey responses is important for the success of surveys under certain conditions. When the survey poses some risks for participants, promises of confidentiality may improve cooperation. Promises of confidentiality are also important to allow respondents to feel comfortable providing answers, especially to sensitive questions. When a survey asks especially sensitive questions, respondents may be more willing to share their thoughts if they know their responses are protected. Some participants would be reluctant to discuss attitudes and opinions on such topics as race, politics, and religion unless they believed their responses could not be identified to them.

Survey research organizations have policies and practices that support confidentiality and use a number of methods to protect confidentiality of survey data. Most organizations require staff members to sign forms stating they will keep the survey data confidential and not reveal any identifiable information outside the survey organization. Survey organizations have elaborate procedures and policies to protect data stored on their computers, particularly data stored on computers that are connected to public computer networks such as the Internet. In some surveys with especially large samples—for example, those conducted by the U.S. Census Bureau—the geographical identifiers could possibly identify respondents. To prevent disclosure of confidential information in these surveys, organizations use a variety of sophisticated data suppression techniques. Because of the multiple data protection methods, survey researchers have a strong record of protecting data integrity and confidentiality.

However, survey data have no clearly defined legal protections that would protect from court subpoenas and possibly other attempts to acquire confidential survey data through the legal system. Fortunately, acquiring identified survey data through legal processes requires substantial effort and is not often successful.

A few exceptions are available to protect survey data legally, but these do not cover most survey research. The U.S. Census Bureau can protect survey data when it collects data under Title 13. This legal protection is especially important for the decennial census, but other surveys are covered by it. Recently, a new confidentiality law—the Confidential Information Protection and Statistical Efficiency Act (CIPSEA)—was enacted to protect data collected by the three federal statistical agencies. The law provides strong confidentiality protections for data collected under it and permits the sharing of the data across the agencies.

Researchers who collect sensitive survey data can apply for Certificates of Confidentiality provided by the National Institutes of Health. The certificate protects the privacy of research subjects such that the investigators and institutions collecting data cannot be compelled to release information that could be used to identify subjects with a research project. The Certification of Confidentiality states that researchers may not be compelled in any federal, state, or local civil, criminal, administrative, legislative, or other proceedings to identify them by name or other identifying characteristic. However, some skepticism exists about whether this protection would survive a serious legal challenge.

The rules on privacy and confidentiality appear to be changing with the widespread use of computer networks and the analysis large scale databases. Yet, survey researchers and survey participants still expect that survey data will remain confidential and protected. The long-term success of the survey industry in protecting its data is important to the profession's overall success.

John Kennedy

See also Anonymity; Cell Suppression; Certificate of Confidentiality; Ethical Principles; Sensitive Topics

Further Readings

National Institutes of Health, Office of Extramural Research. (n.d.). *Certificates of Confidentiality kiosk.* Retrieved January 4, 2007, from http://grants1.nih.gov/grants/policy/coc/index.htm

Singer, E., Van Hoewyk, J., & Neugebauer, R. J. (2003). Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 Census. *Public Opinion Quarterly*, *67*, 368–384.

CONSENT FORM

In survey research, consent forms typically are used to gain the permission of a parent or guardian who has the legal authorization to give permission for someone in her or his charge to participate in a survey. However, in some studies an adult will be asked to sign a consent form about her or his own agreement to participate in a survey.

Consent forms are most commonly used in surveys of youth populations, regardless of survey mode. Federal regulations protecting human subjects (45 CFR 46), accompanying state or local regulations, and many institutional review boards (IRBs) hold that a youth cannot legally agree to complete a survey (provide consent for herself or himself) until he or she is 18 years of age. As a result, signed or written permission from a parent or legal guardian usually is required prior to the youth or child participating in a survey. This permission is obtained by providing a written permission form, called a "consent form," to parents and having a parent or guardian return it with his or her signature giving the child permission to participate in the survey. Consent forms document that youth have permission to participate in the survey and help ensure that parents or guardians have enough information about the survey to make a decision about whether the youth can participate.

Consent forms also can be required for surveys of adult populations; a key difference with adult populations is that the adult respondent is asked to sign the consent form documenting that she or he has enough information about the survey to make an informed decision to participate.

Under federal human subjects protection regulations (45 CFR 46.116(a)), consent forms usually must include the following elements (individual institutional review boards may require additional elements):

1. An explanation of the purposes of the survey, the expected length of the survey, and a description of the procedures to be followed
2. A description of any reasonably foreseeable risks or potential harm that could occur if the respondent participates in the survey
3. A description of any benefits to the respondent or to others that may be expected from the survey or that may be provided directly by the researchers
4. A statement describing the extent to which confidentiality of any answers or data identifying the respondent will be maintained by researchers
5. Details about whom to contact for answers to questions about the survey and about respondents' rights, and information about whom to contact if participation in the survey results in any harm to the respondent, and
6. A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the respondent is otherwise entitled, and a statement that the respondent may terminate participation at any time without any penalty

Although consent forms usually are required for surveys of youth populations, federal regulations and IRBs often provide some flexibility for surveys of adult populations. For adult populations, participation in surveys rarely puts respondents at more than the minimal risks of everyday life. Moreover, depending on the mode of a survey, documentation of consent may not be feasible and may harm surveys by significantly reducing response rates. Finally, some surveys of sensitive behavior rely on anonymity to increase the likelihood that respondents answer questions honestly; for these surveys, a signed consent form actually serves as the only link between a respondent and his or her answers, thus making anonymity impossible and providing a possible threat to confidentiality. As a result, IRBs often waive requirements of a consent form and a signature for surveys with adult populations and allow the informed consent process to occur informally as part of the survey itself. However, key elements of consent can be provided to respondents in a concise way at the beginning of a survey—in the introductory script in a telephone interview, in a cover letter for a self-administered survey, or on the introductory screen in a Web survey.

Matthew Courser

See also Informed Consent; Institutional Review Board (IRB); Protection of Human Subjects

Further Readings

American Association for Public Opinion Research. (n.d.). *Institutional review boards.* Retrieved March 24, 2008, from http://www.aapor.org/institutionalreviewboards

U.S. Department of Health and Human Services. (2005). *Code of Federal Regulations, Title 45 Public Welfare and Part 46 Protection of Human Subjects.* Retrieved March 17, 2008, from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

CONSTANT

The term *constant* simply refers to something that is not variable. In statistics, and survey research in particular, responses are typically described as random variables, roughly meaning that the responses cannot be predicted with certainty. For example, when people are asked whether they approve or disapprove of a particular political leader, typically there is uncertainty about what the response will be. As another example, in a survey regarding whether individuals approve or disapprove of the death penalty, responses are not constant simply because some individuals will approve and others will not.

Although at some level, the difference between a constant and a random variable is clear, the distinction between the two often becomes blurred. Consider, for example, the population mean, μ. That is, μ is the average of all individuals of interest in a particular survey if they could be measured. The so-called frequentist approach to statistical problems views μ as a constant. It is some fixed but unknown value. However, an alternative view, reflected by a Bayesian approach to statistics, does not view μ as a constant, but rather as a quantity that has some distribution. The distribution might reflect prior beliefs about the likelihood that μ has some particular value.

As another example, p might represent the probability that an individual responds "Yes" when asked if he or she is happily married. In some sense this is a constant: at a particular moment in time one could view p as fixed among all married couples. Simultaneously, p could be viewed as a random variable, either in the sense of prior beliefs held by the investigator or perhaps as varying over time.

Another general context in which the notion of constant plays a fundamental role has to do with assumptions made when analyzing data. Often it is assumed that certain features of the data are constant in order to simplify technical issues. Perhaps the best-known example is *homoscedasticity*. This refers to the frequently made assumption that the variance among groups of individuals is constant. In regression, constant variance means that when trying to predict Y based on some variable X, the (conditional) variance of Y, given X, does not vary. So, for example, if X is amount of solar radiation associated with a particular geographic region, and Y indicates breast cancer rates, constant variance means that the variance of Y does not differ among the geographic regions that are of interest.

Rand R. Wilcox

See also Variable

CONSTRUCT

In the context of survey research, a construct is the abstract idea, underlying theme, or subject matter that one wishes to measure using survey questions. Some constructs are relatively simple (like political party affiliation) and can be measured using only one or a few questions, while other constructs are more complex (such as employee satisfaction) and may require a whole battery of questions to fully operationalize the construct to suit the end user's needs. Complex constructs contain multiple dimensions or facets that are bound together by some commonality that, as a whole, compose the construct. Without clearly conceptualizing the construct's dimensions and the common theme binding the dimensions together, the survey developer runs the risk of either creating a set of questions that does not measure all of what is intended or creating a set of questions that measures dimensions of an unintended construct.

Before question writing or compilation begins, the construct should be carefully considered and its relevant dimensions defined. As a cohesive set, the dimensions of a construct define the construct. Some constructs are relatively simple and do not have many dimensions. For example, the construct of political party identification is relatively simple and may require only a question or two in order to adequately encompass its dimensions. For years, the General Social Survey has asked the question *Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?* with response options ranging from "Strong Democrat" to "Strong Republican." That one question adequately covers political party affiliation and strength of party identification, which are two relevant dimensions of the construct.

However, the broader a construct, the more dimensions it generally contains. For example, the construct "employee satisfaction" is a broad construct with many dimensions. Simply asking employees the question *How satisfied are you with your job?* is far from adequate. The construct of employee satisfaction has many dimensions that may include the company's culture and values, organizational leadership style, pay structure, working conditions, opportunities for advancement, long-term plans, and training. Each of these dimensions might be further broken down into smaller subdimensions that are more easily operationalized into separate questions.

If a construct is the abstract subject matter to be measured, operationalization is the concrete and measurable expression of the dimensions of that idea in the form of a question or questions. "Working conditions" is a dimension within the construct of employee satisfaction. This dimension of employee satisfaction could be examined using multiple questions dealing with topics ranging from the comfort of the desk chairs to the number of hours employees are expected to work in a normal week. It is the responsibility of those creating the questionnaire to determine the construct dimensions that are most important and operationalize accordingly. Various statistical methods such as factor analysis can help determine the centrality of operationalized questions to the construct.

Dennis Dew

See also Construct Validity; Questionnaire Design; Reliability; Validity

Further Readings

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Sudman, S., & Bradburn, N. M. (2004). *Asking questions: The definitive guide to questionnaire design—For market research, political polls, and social and health questionnaires.* San Francisco: Jossey-Bass.

Construct Validity

In survey research, construct validity addresses the issue of how well whatever is purported to be measured actually has been measured. That is, merely because a researcher claims that a survey has measured presidential approval, fear of crime, belief in extraterrestrial life, or any of a host of other social constructs does not mean that the measures have yielded reliable or valid data. Thus, it does not mean the constructs claimed to be measured by the researcher actually are the ones that have been measured.

In most cases, for survey measures to have high construct validity they also should have good "face validity." Face validity is a commonsensical notion that something should at least appear on the surface (or "at face value") to be measuring what it purports to measure. For example, a survey item that is supposed to be measuring presidential approval that asks, *How well is the country being run by the current administration*? has only some face validity and not much construct validity. Its face validity and thus its construct validity would be enhanced by adding the name of the president into the question. Otherwise it is a stretch to claim that the original wording is measuring the president's approval. One reason for this is that there could be many other members of "the current administration" other than the president who are affecting the answers being given by respondents.

The single best way to think about the likely construct validity of a survey variable is to see the full wording, formatting, and the location within the questionnaire of the question or questions that were used to gather data on the construct. In this way one can exercise informed judgment on whether or not the question is likely to have high construct validity. In exercising this judgment, one should also consider how the question was administered to the respondents and if there is anything about the respondents themselves that would make it unlikely for them to answer accurately. Unfortunately, too few consumers of survey results have access to this detailed type of information or take the time to think critically about this. This applies to too many journalists who disseminate survey information without giving adequate thought to whether or not it is likely to have solid construct validity.

For researchers and others who have a greater need to judge the construct validity of variables on the basis of empirical evidence, there are a number of statistical analyses that can (and should) be performed. The simpler of these analyses is to investigate whether answers given by various demographic groups are within reasonable expectations. For example, if it is reasonable to expect gender differences, are those gender differences actually observed in the data? Additional, correlational analyses should be conducted to determine if the variables of interest correlate with other variables they should relate to. For example, if a Democrat is president, do respondents who are strong Republicans give considerably lower approval ratings than respondents who are strong Democrats? A final consideration: variables that are created from multiple survey items, such as scaled variables, should be tested to learn if they have strong internal consistency using procedures such as factor analyses and calculating Cronbach's alpha. If they do not, then one should suspect their construct validity.

Paul J. Lavrakas

See also Construct; Cronbach's Alpha; Interviewer-Related Error; Measurement Error; Mode-Related Error; Questionnaire-Related Error; Reliability; Respondent-Related Error; Validity

Further Readings

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *American Psychologist*, *52*, 281–302.

Campbell, D. T., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *3*, 635–694.

CONSUMER SENTIMENT INDEX

The Consumer Sentiment Index is a measure of consumer confidence in the United States that has been measured and reported by the University of Michigan every month, starting in the early 1950s.

Consumer sentiment, which often is called "consumer confidence," is cited by government officials, business executives, the media, and by ordinary citizens to describe national economic conditions. It has become so much a part of the national economic dialogue that many people think that consumer confidence has a specific and widely agreed-upon definition. Nonetheless, the definition of consumer confidence has remained elusive, since the confidence of consumers can never be directly observed; it is only the behavior of consumers that can be observed. Interest in consumer confidence is thus defined by an interest in the economic behavior of consumers. It is the consumer who determines whether the economy moves toward expansion and growth or toward contraction and recession. Indeed, consumer spending and residential investment account for three quarters of all spending in the U.S. domestic economy, and consumers invest more in homes, vehicles, and other durable goods than business firms invest in new structures and equipment.

The usefulness of measures of consumer sentiment as leading economic indicators has garnered worldwide recognition and is now measured by countries in all six inhabited continents. The countries include Argentina, Austria, Australia, Belgium, Bulgaria, Brazil, Canada, China, Cyprus, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hong Kong, Hungary, Indonesia, Ireland, Italy, Jamaica, Japan, Korea, Lithuania, Luxembourg, Latvia, Malaysia, the Netherlands, Mexico, Norway, New Zealand, Poland, Portugal, Romania, Russia, Spain, the Slovak Republic, Slovenia, South Africa, Sweden, Switzerland, Taiwan, Thailand, Turkey, the United Kingdom, and the United States. In addition, there are a large number of other Central and South American countries that have measured consumer confidence sporadically without the establishment of a consistent time series.

Early Development

An economic behavior research program at the University of Michigan began as part of the post–World War II planning process. Its agenda was focused on understanding the role of the consumer in the transition from a wartime economy to what all hoped would be a new era of peace and prosperity. The primary purpose of the first survey in 1946 was to collect in-person data on household assets and debts. The sponsor of the survey, the Federal Reserve Board, initially had little interest in the attitudes and expectations of consumers. Their goal was a financial balance sheet, the hard currency of economic life, not the soft data of consumer sentiment. George Katona, the founder of the survey program, convinced the sponsor that few respondents would be willing to cooperate if the first question asked was, *We are interested in knowing the amount of your income and assets. First, how much do you have in your savings account?* Instead, sound survey methodology required that other, more general, and less threatening questions were first needed to build respondent rapport and to establish a sense of trust and confidence with the respondents.

Katona devised a conversational interview that introduced each new area of interest with questions that first elicited general opinions before asking the detailed questions on dollar amounts. Although the sponsor was convinced that such attitudinal questions were needed for methodological reasons, Katona was told that he did not need to report any of these results since the Federal Reserve had no interest in the attitudinal findings. Ultimately, the Federal Reserve Board, as well as many others, became as interested in the findings on consumers' expectations as on consumers' balance sheets. Although the first measures of consumer expectations may seem serendipitous, it was in reality no happenstance. Katona had clear and unmistakable intentions and seized this opportunity to give life to an innovative research agenda. Katona had long been interested in the interaction of economic and psychological factors, what he termed "the human factor" in economic affairs. When Katona advocated his theory of behavioral economics, few economists listened; 50 years later behavioral economics is at the center of new theoretical developments.

When the sentiment measure was first developed in the late 1940s, it was intended to be a means to directly incorporate empirical measures of income expectations into models of spending and saving behavior. Katona summarized his views by saying that consumer spending depends on both consumers' ability and willingness to buy. By *spending*, he meant discretionary purchases; *by ability*, he meant the current income of consumers; and by *willingness*, he meant consumers' assessments of their future income prospects. Katona hypothesized that spending would increase when people became optimistic, and precautionary saving would rise when they became pessimistic.

Consumer confidence was originally conceptualized as a broad measure of expected changes in income. It was not simply the expected size of a consumer's future income, but the certainty or uncertainty that was attached to those expectations. Thus, an important component of the definition of consumer confidence was that it encompassed both the expected level as well as the expected variance of income. To recognize this dual criterion, Katona defined the *dimension* of consumer confidence as ranging from optimism and confidence to pessimism and uncertainty.

Moreover, Katona argued that consumer confidence has affective as well as cognitive dimensions. Indeed, it was this recognition that led Katona to change the name of the index from "Consumer Confidence" to "Consumer Sentiment." Katona recognized that few consumers thought of inflation or unemployment, for example, without making evaluative judgments. The affective components of economic attitudes and expectations are what serve to integrate diverse pieces of economic information. Moreover, it is the affective component that enables waves of optimism or pessimism to sweep across the population with great speed.

The University of Michigan's Index of Consumer Sentiment was formed at the start of the 1950s when sufficient time-series data had been collected. The index is based on the responses to five questions—two questions on personal finances, two on the outlook for the economy, and one question on buying conditions for durables:

1. *We are interested in how people are getting along financially these days. Would you say that you (and your family) are better off or worse off financially than you were a year ago?*
2. *Now looking ahead—do you think that a year from now you (and your family) will be better off financially, or worse off, or just about the same as now?*

3. *Now turning to business conditions in the country as a whole—do you think that during the next twelve months we'll have good times financially, or bad times, or what?*
4. *Looking ahead, which would you say is more likely—that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have periods of widespread unemployment or depression, or what?*
5. *About the big things people buy for their homes—such as furniture, a refrigerator, stove, television, and things like that. Generally speaking, do you think now is a good or a bad time for people to buy major household items?*

While Katona would have preferred to report on the detailed findings from the surveys, he recognized that a summary index was needed for both the ease of dissemination as well as empirical testing. It is inherently difficult to summarize the diverse implications for all forms of consumer spending in a single index, and there was never an attempt to do so. Indeed, the Michigan surveys include a large range of additional questions. The questions range from income, unemployment, interest rates, and inflation expectations to what respondents think are the most important recent changes in economic conditions, measures about buying conditions for a variety of products, attitudes toward savings and debt, holdings of various assets, and many other topics.

Adaptation to Change

In the late 1940s, most consumers viewed all aspects of the economy through the single dimension of how it affected their jobs and income prospects. In the 21st century, while job and income prospects are still important, there are many other aspects of the economy that are just as important to consumers. For example, consumer expectations for interest rates, inflation, stock prices, home values, taxes, pension and health care entitlements as well as jobs and incomes have moved independently, and often in opposite directions. Furthermore, consumers are now more likely to make distinctions between the near- and longer-term prospects for inflation and stock prices as well as between near- and longer-term job and income prospects. Moreover, consumers have also recognized the importance of the global economy in determining wage and job prospects as well as determining the prices of products sold on Main Street and financial assets sold on Wall Street.

Demographic shifts also influence the measurement of confidence. The retirement of the baby boom generation will reduce their concerns about adverse developments in domestic labor markets in comparison to their heightened dependence on inflation-adjusted returns on their retirement savings. The impact of globalization on financial markets is far greater and nearly instantaneous compared with its influence on labor markets. In addition, older consumers have different spending priorities, and it can be expected that the importance of low import prices for durable goods will fade in comparison to the provisions for health care and other services to the elderly. More generally, the trend toward the increase in purchases of services compared with the purchase of durable goods requires a new conceptualization of consumer confidence.

As a result, the measurement of consumer confidence will likely become even more challenging as it continues to expand into a broader and more complex assessment of economic prospects. Indeed, the economic environment may have become too diverse, and consumers too sophisticated, for any single index to accurately and unambiguously describe consumers as either optimistic or pessimistic. It may be true that no single index can be devised to accurately predict all types of expenditures for all types of consumers at all times. The most accurate models of consumer behavior will relate specific spending and saving behaviors to specific expectations. Nonetheless, there is still a need for an overall index of consumer sentiment that broadly summarizes trends, just as there is a need for aggregate statistics such as the gross domestic product (GDP).

Along with the growing sophistication among consumers, there is a growing demand for more precise measures of expectations. As expectations have become more central components of economic models, the theoretical specifications of the desired measures have become more exacting. Economists generally favor probability measures, while psychologists generally favor verbal likelihood questions. Numeric probability scales are assumed to allow the comparability of responses among different people, across situations, and over time. The simple formulations of verbal likelihood scales, in contrast, are presumed to be answerable by nearly everyone, even by those with limited information or computational skills. The Michigan surveys now incorporate both types of measures.

The strength of household surveys is that they are based on the premise that the description and prediction of consumer behavior represent the best means to foster advances in theory. While there is nothing more useful than good theory, there is nothing more productive in generating theoretical advances than good data. To this end, the Michigan surveys have always stressed the substance of the research rather than the format of the questions or the components of the sentiment index. The more rapid changes that may accompany an aging population and the globalization of the economy are seen as an opportunity for scientific advancement. Consumer confidence will still be part of popular culture, still be thought to have a specific and widely agreed-upon definition, and still be an unobserved variable that is defined by the evolving economic behavior of consumers.

Sample Design

The monthly survey is based on a representative sample of all adult men and women living in households in the coterminous United States (48 states plus the District of Columbia). A one-stage list-assisted random-digit design is used to select a probability sample of all telephone households; within each household, probability methods are used to select one adult as the designated respondent. The probability design permits the computation of sampling errors for statistics estimated from the survey data.

The sample is designed to maximize the study of change by incorporating a rotating panel in the sample design. An independent cross-section sample of households is drawn each month. The respondents chosen in this drawing are then reinterviewed six months later. A rotating panel design results, with the total of 500 monthly interviews made up of 60% new respondents and 40% being interviewed for the second time. The rotating panel design has several distinct advantages. This design provides for the regular assessment of change in expectations and behavior both at the aggregate and at the individual level. The rotating panel design also permits a wide range of research strategies made possible by repeated measurements. In addition, pooling the independent cross-section samples permits the accumulation of as large a case count as needed to achieve acceptable standard errors for critical analysis variables.

Richard Curtin

See also Cross-Sectional Survey Design; List-Assisted Sampling; Random-Digit Dialing (RDD); Rotating Panel Design

Further Readings

Curtin, R. (1983). Curtin on Katona. In H. W. Spiegel & W. J. Samuels (Eds.), *Contemporary economists in perspective* (Vol. 1, pp. 495–522). New York: Jai Press.
Curtin, R. (2004). Psychology and macroeconomics. In J. S. House, F. T. Juster, R. L. Kahn, H. Schuman, & E. Singer (Eds.), *A telescope on society* (pp. 131–155). Ann Arbor: University of Michigan Press.
Katona, G. (1951). *Psychological analysis of economic behavior*. New York: McGraw-Hill.
Katona, G. (1960). *The powerful consumer: Psychological studies of the American economy*. New York: McGraw-Hill.
Katona, G. (1964). *The mass consumption society*. New York: McGraw-Hill.
Katona, G. (1975). *Psychological economics*. New York: Elsevier.

CONTACTABILITY

The ease or difficulty with which a sampled respondent can be contacted by a survey organization is referred to as her or his “contactability.” It can be expressed as a quantity (or “contact propensity”) and ranges from 0.0 to 1.0, with 0.0 meaning it is impossible to contact the respondent and 1.0 meaning it is certain that the respondent will be contacted.

Contactability will vary by the mode that is used to attempt to contact a respondent in order to recruit her or his cooperation and/or gather data. Contactability also will vary according to the effort a survey organization expends to reach the respondent and what days and times these contact attempts are tried.

For example, take the case of young adult males, ages 18 to 24 years, who are among the hardest of demographic groups for survey organizations to make contact with. The mode of contact that is used will affect the contactability of this cohort, as they are far less likely to be contacted via a traditional random-digit dialed (RDD) landline survey. If the telephone mode is used, then researchers trying to contact this cohort also need to sample cell phone numbers, as nearly one third of these adults in the United States were “cell phone only” in 2007 and their proportion is growing each year. If the mode of contact is the postal service (mail), this young adult male cohort also will have relatively lower contactability, as they are likely to move from address to address more than other demographic groups.

The number of days, which days of the week, and what times of day a survey organization uses its interviewers (telephone or in-person) to make contact with respondents also will affect the contactability of respondents. In the case of the young adult cohort, fielding the survey for only a few days (e.g., a weekend poll, Friday through Sunday) will greatly lower the contactability of this cohort, especially if no late evening hours are included.

In a telephone survey, contactability also will vary by whether or not the survey organization sends out some form of name identifier to be shown on caller ID or on the privacy manager devices that many households use to decide whether or not to answer their incoming calls. (Yet, even if the survey organization's name is displayed on such a device it will not help raise contactability unless the respondents know the name and think positively toward it.) Leaving a message on an answering machine when it is first encountered at a household is thought to aid contactability, assuming the message is a persuasive one, given that many household use these machines to screen their incoming calls.

Low levels of contactability within a sample will lead to higher nonresponse due to noncontact. Thus, it behooves researchers to think explicitly about cost-effective ways to increase the contactability of their sampled respondents.

Paul J. Lavrakas

See also Calling Rules; Callbacks; Contact Rate; Contacts; Mode of Data Collection; Noncontacts; Nonresponse

CONTACT RATE

Contact rate measures the proportion of eligible cases in the sampling pool in which a member of a sampled household was contacted—that is, reached by an interviewer (in telephone and in-person surveys) or received the survey request (in the case of mail and Internet surveys). Contact rates can be computed for all surveys, regardless of the mode in which the data are gathered. The contact rate is a survey outcome rate that can be cited in survey reports and in research literature. Although no single rate or number can reflect the total quality of a survey, contact rates (along with survey response rates, survey cooperation rates, and survey refusal rates) are one of the most common outcome tools that researchers use to evaluate survey quality.

Both household-level and respondent-level contact rates can be computed for a survey by using the final sample dispositions. In the former case, the household-level contact rate reflects the proportion of cases in which any sort of contact was made with a person at a household, including cases in which contact was made with eligible respondents. The respondent-level contact rate is similar, with the exception that it reflects only the proportion of contacts with known survey respondents. Researchers typically compute 1 of 3 standard contact rates.

Contact Rate 1

The numerator of this rate is comprised of all of the kinds of contacts (e.g. completion, refusals, language barrier, and so on) a survey or interviewer (depending on the mode) might make with a person in a sampled household or unit (or with the respondent, if a respondent-level contact rate is being computed). The denominator includes all known eligible cases and all cases of indeterminate eligibility. As such, this rate is the most conservative contact rate.

Contact Rate 2

As before, the numerator of this rate is comprised of all of the kinds of contacts a survey or interviewer (depending on the mode) might make with a person in a sampled household or unit (or with the respondent, if a respondent-level contact rate is being computed). However, the denominator of this rate includes all known eligible cases and a proportion of the cases of indeterminate eligibility that is based on the researcher's best estimate of how many of the cases of indeterminate eligibility actually are eligible.

Contact Rate 3

As with Contact Rates 1 and 2, the numerator of this rate is comprised of all of the kinds of contacts a survey or interviewer (depending on the mode) might make with a person in a sampled household or unit (or with the respondent, if a respondent-level contact rate is being computed). The denominator of this rate includes only the known eligible cases. As a result, Contact Rate 3 is the most liberal contact rate.

Matthew Courser

See also Contacts; Cooperation Rate; Eligibility; Final Dispositions; Refusal Rate; Response Rates; Sampling Pool; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

CONTACTS

Contacts are a broad set of survey dispositions that are used with all surveys (telephone, in-person, Internet, and mail), regardless of mode. The set of contact dispositions includes all the kinds of contacts a survey or interviewer (depending on the mode) might make with a person or sampled household or unit.

Many of the most common types of contacts occur in all surveys, regardless of the mode in which they are conducted. These include completed interviews, partial interviews, refusals, and breakoffs. Other, less common types of contacts include cases in which contact is made with a respondent or sampled unit or household but an interview is never started because the sampled respondent is physically or mentally unable to participate, or an interviewer is told the respondent is unavailable to complete the questionnaire during the entire field period. Contacts also include cases involving language barriers (with a telephone or in-person survey) and literacy issues relating to respondents not being able to read and understand the questionnaire, in the case of mail and Internet surveys. A final type of contact occurs when it is determined that the person or household is ineligible for the survey. Of note, in many cases in mail and Internet surveys, the researcher has no idea whether or not contact ever was made with anyone at the sampling unit.

Contacts are used for computing contact rates for surveys. A contact rate measures the proportion of all cases in the sampling pool in which a member of a sampled household was reached by an interviewer (in telephone and in-person surveys) or received the survey request (in the case of mail and Internet surveys).

Matthew Courser

See also Completion Rate; Contact Rate; Language Barrier; Partial Completion; Refusal; Refusal Rate

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

CONTENT ANALYSIS

As it relates to survey research, content analysis is a research method that is applied to the verbatim responses given to open-ended questions in order to code those answers into a meaningful set of categories that lend themselves to further quantitative statistical analysis. In the words of Bernard Berelson, one of the early scholars explaining this method, "Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication." By coding these verbatim responses into a relatively small set of meaningful categories, survey researchers can create new variables in their survey data sets to use in their analyses.

Example of Content Analysis in Survey Research

Imagine a questionnaire that asks respondents, *What is the biggest problem facing the nation today?* Some of the answers that respondents have given to this open-ended question are shown in Figure 1 (along with the spelling and grammar mistakes made by telephone interviewers).

For a survey researcher to be able to analyze the "biggest problem" question, these verbatim answers must be coded into a relatively small and meaningful set of categories. Using the verbatims in Figure 1, a plausible set of categories could be as follows:

President Bush; His administration and its policies

The Republican Congress

ILLWILLED AND MALICE DRIVEN BY IGNORANCE
LOSS OF CIVIL LIBERTIES
THE PRICES OF HOUSES ARE TOO HIGH
TRUST
MORAL DECLINE
WELFAIR
THE PRESIDENT
REPUBLICAN ADMISTRATION
THE BUDGET DIFFCIET
NATIONAL SECURITY
ILLIGAL IMMIGRATION
A PRESIDENT WHO DOESNT UNDERSTAND INTERNATIONAL POLITICS
NATIONS DEBT
THE LACK OF PUTTING GOD FIRST IN OUR LIVES
KILLING
GAS PRICES
UNEMPLOYMENT
ILLEGAL ALIENS
MORAL DECLINE
ECONOMY AND WAR IN IRAQ
BAD GOVERNMENT
APATHEY NOBODY CARING ABOUT ANYTHING
NEIGHBORHOOD SECURITY
TAKING GOD EVERYTHING
GEORGE BUSH
PRSEDENT THAT A LEIR
GEORGE BUSH
TO MANY PEOPLE TO MANY COMEING INTO THE US
PRESENT ADMINISTRATION
OUR ADMINISTRATION OR OUR GOVERNMENT OUR CURRENT POLICIES
A DISHONEST GOVERNMENT
CORRUPTION IN OUR GOGERNMENT
GOVERNMENT POLICY REGARDING IRAQ AND TERROR ACTIVITY AROUNFD THE WORLD
GREED
REPUPLICANS CONTROLLING LEGISLATIVE AND EXECUTIVE BRANCHES AT THE NATIONAL LEVEL
AFFORDABLE HOUSING

Figure 1 Examples of answers given to open-ended question, "What is the biggest problem facing the nation today?"

Honesty in government

Immigration; Illegal aliens

Moral decline

Housing

War in Iraq

National security; Terrorism

Misc. Other

Coders need to be carefully trained and regularly monitored to apply these categories reliably to each verbatim answer and thereby assign a numerical value to that answer. In this example, a new coded variable would be created that ranges in value from 1 to 9. This variable then could be analyzed via cross-tabulation or other statistical procedures to learn, for example, whether certain demographic characteristics of the respondents (e.g., age, gender, and race) are related to the answers given to the open-ended question.

Content analysis can also be performed by computer software programs. Again, the researchers need to devise a reliable coding scheme in order for the end product to be reliable. For many researchers, the limitations of what software can accomplish are offset by the lower costs of doing the content coding with software compared to the much higher cost of doing it with human coders. However, many content coding solutions will be beyond the capacity of current computer software to apply reliably, and in those instances human coders will need to be utilized.

Analytic Considerations

A general rule of thumb that many survey researchers have found in doing content analyses of open-ended answers is to code as many as three new variables for each open-ended question. For example, if the open-ended question is Q21 in the questionnaire, then the three new variables might be named Q21CAT1, Q21CAT2, and Q21CAT3. This follows from experience that indicates that nearly all respondents will give at least one answer to an open-ended question (since most of these open-ended questions do not ask for only one answer). Many respondents will give two answers, and enough will give three answers to justify coding up to three answers from each respondent. When this approach is used, the researcher also may want to create other new dichotomous (dummy) variables coded "0" or "1" to indicate whether each respondent did or did not mention a certain answer category. Thus, for the earlier example using the "biggest problem" question, new dichotomous variables could be created for each category (BUSH, CONGRESS, HONESTY, IMMIGRATION, etc.). For each of these new variables, the respondent would be assigned the value of "0" if she or he did not mention this category in the open-ended verbatim response and "1" if this category was mentioned.

Paul J. Lavrakas

See also Content Analysis; Open-Ended Question; Verbatim Responses

Further Readings

Berelson, B. (1954). Content analysis. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 1, pp. 488–522). Reading, MA: Addison-Wesley.

Franzosi, R. (2003). *From words to numbers.* Cambridge, UK: Cambridge University Press.

Krippendorf, K. (1980). *Content analysis: An introduction to its methodology.* Beverly Hills, CA: Sage.

Weber, R. P. (1990). *Basic content analysis.* Newbury Park, CA: Sage.

Context Effect

The term *context effect* refers to a process in which prior questions affect responses to later questions in surveys. Any survey that contains multiple questions is susceptible to context effects. Context effects have the potential to bias the thinking and answers of survey respondents, which reduces the accuracy of answers and increases the error in survey measurement. Psychologists refer to context effects as the general effect of *priming*. Priming occurs when the previous activation of one type of information in active memory affects the processing of subsequent related information. For example, the prior presentation of the word *doctor* reduces the time it takes to subsequently recognize the word *nurse* in comparison to an unrelated word. This priming effect is thought to occur because the activation of one concept spreads and activates related concepts in the brain. Similarly, for example, attempting to remember a list of words that all relate to "bed" (i.e., *sleep, pillow*, etc.) increases the likelihood that a person will falsely remember that the related word was present in the list during recall. In both cases, the previous context consistently primes, or biases, thinking in a certain direction by increasing the saliency of that information.

Context effects are most noticeable in attitude surveys. These contexts effects may occur (a) within a question, and (b) between questions (also referred to as "question order effects"). An example of a within-question context effect is how the label *anti-abortion* instead of *pro-choice* affects attitudes toward abortion. The wording choice leads the respondent to frame a question in a certain way or increases the saliency and importance of some information over other information within a question. A between-question context effect occurs, for example, when previous questions regarding attitudes toward an ongoing war influence a subsequent question regarding presidential performance. Question order effects are evident in the fact that answers to questions on related themes are more similar and consistent when the questions are asked in a group than when these questions are separated and scattered throughout a questionnaire. Effects of question order are also evident when questions regarding a negative life event lead to more negative attitudes for subsequent questions regarding present feelings.

It is possible to control for context effects by counterbalancing question order across several versions of a survey. However, due to cost concerns, this option is rarely feasible to implement properly. It is unavoidable that the wording of survey questions frames and defines issues for survey respondents in ways that affect responses. Questions will be interpreted by respondents within the context of the entire questionnaire, previous questions, and the wording of the present question. Given that these processes are unavoidable and cannot be eliminated, survey designers must at least be aware of the possible effects of context and thereby try to design questionnaires in order to minimize their effect. Question construction must balance the positive impact of greater question detail on retrieval performance with the negative effects leading respondents toward certain responses because of greater detail.

It should be noted that although awareness of possible context effects is advisable, there is actually little evidence that context effects have a great impact on most overall survey results. The percentage of questions in any survey affected by context effects in any significant way tends to be around only 5%. Thus, even though a few particular items may be affected by prior information, context effects rarely appear to alter survey answers away from a respondent's "true" answers to any great extent across whole surveys.

Gregory G. Holyk

See also Cognitive Aspects of Survey Methodology (CASM); Measurement Error; Priming; Question Order Effects; Saliency

Further Readings

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, pp. 82–90.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys.* New York: Academic Press.

Contingency Question

Questions that are limited to a subset of respondents for whom they are relevant are called "contingency questions." Relevancy is sometimes based on a respondent characteristic such as gender or age. For example, it is typical to ask only women of childbearing age if they are currently pregnant; conversely, only men are asked if they have ever have had a prostate cancer screening examination. Other times, questions are asked only of those that engage in a certain activity or hold a certain opinion about an issue.

A question that determines if a contingency question is asked is called a "filter," "skip," or "branching" question. In the research literature, the terms *filter question* and *contingency question* are sometimes used synonymously. However, in practice, the latter is dependent, or contingent, on the response to the former. Filter questions help route respondents through the questionnaire by skipping them over questions that are not relevant. Questionnaire "pathing" can be simple, as when one filter question determines receipt of one contingency question. Complexity is increased when responses to a series of filter questions are used to determine if a respondent gets one or more contingency questions.

Filter and contingency questions can be deployed in any data collection mode. In certain modes (Web, computer-assisted telephone interview [CATI], computer-assisted personal interviewing [CAPI], or computer-assisted self-interviewing [CASI]), the determination of who receives a contingency question can be programmed electronically. Once respondent characteristics are pre-loaded, survey programs will automatically skip contingency questions that would otherwise have required asking one or more filter questions. For example, respondents who are known to be male would automatically skip questions contingent on being female without first being asked a filter question about gender.

Contingency questions are not required on survey instruments; however, their use, in conjunction with filter questions, can reduce overall burden by asking respondents only those questions that are relevant. In the absence of filter questions, "Not Applicable" should be added as a response category for items relevant to only a subset of respondents. In the absence of an explicit Not Applicable option, respondents for whom inapplicable questions are asked may respond with a "Don't Know" or "Refuse." This could be interpreted erroneously as missing data.

Survey researchers should be cognizant of the fact that some respondents may purposely answer filter questions in a way that will result in skipping contingency questions. This may occur when respondents lose interest in the survey, whether it is due to fatigue, boredom, or lack of topic saliency, and when they can too easily anticipate how a particular answer to a filter question will skip them out of another question or series of questions. This can lower data quality, as the result would be undetected missing data on items for which a respondent was actually eligible.

Kirsten Barrett

See also Missing Data; Questionnaire Design; Respondent Burden

Further Readings

Babbie, E. R. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Dillman, D., Carley-Baxter, L., & Jackson, A. (1999). *Skip pattern compliance in three test forms: A theoretical and empirical evaluation*. Technical report no. 99-01. Pullman: Washington State University, Social and Economic Sciences Research Center.

Contingency Table

A contingency table (or *cross-tabulation*) is an effective way to show the joined distribution of two variables, that is to say, the distribution of one variable within the different categories of another. Data in the table are organized in rows and columns. Each row corresponds to one category of the first variable (usually considered as the dependent variable), while each column represents one category of the second variable (usually considered as an independent variable). The intersection of a row and a column is called a "cell." Each cell contains the cases that have a certain combination of attributes corresponding to that row and column (see Table 1). Inside each cell a variety of information can be displayed, including (a) the total count of cases in that cell, (b) the row percentage represented by the cell, (c) the column percentage represented by the cell, and (d) the proportion of the total sample of cases represented by that cell.

Generally, a contingency table also contains the sums of the values of each row and column. These sums are called the "marginals" of the table. The sum of column or row marginals corresponds to the sample size or grand total (in the lower right-hand cell of the table).

The product of the number of the rows by the number of the columns is called the "order" of the table (Table 1, for example, is a 2×2 table), while the number of the variables shown in the table represents its dimension.

Table 1 An example of contingency table: Gender by education

		Gender		
		M	*F*	*Total*
Education	*Low*	32	87	119
		26.9%	73.1%	100%
		17.0%	37.5%	28.3%
		7.6%	20.7%	
	High	156	145	301
		51.8%	48.2%	100%
		83.0%	62.5%	71.7%
		37.1%	34.5%	
	Total	188	232	420
		44.8%	55.2%	100%
		100%	100%	

A bivariate contingency table represents the first device the researcher can use in the exploration of the relationship between two variables (including ones that are nominal or ordinal). In order to establish whether the variables are associated or not, however, the researcher has to abandon the raw frequencies in favor of the percentages, because only these allow a proper comparison. One can calculate three types of percentages: (1) row, (2) column, and (3) total percentages. However, not all these percentages are generally reported in the contingency table, as that would be more information than needed in most instances; although they are shown in each cell in Table 1 below the cell count. Which percentages the researcher takes into account depends on the specific research question. However, if the researcher aims at exploring the influence of the variable shown in the columns (considered as independent) on the variable shown in the rows (considered as dependent), she or he should report the column percentages. Therefore, keeping fixed the first category of the dependent variable (in the rows), the researcher will analyze how the values change along the categories of the independent variable (in the columns). If one considers the column percentages in the Table 1 (i.e., the 2nd percentage below the count in each cell) for example, keeping fixed the category "low educated," one can see that females in this sample are significantly more likely to be "less educated" than are males. Of note, if the percentages in a cell are based on too small a number of cases, the results will not be reliable.

Contingency tables with the same number of rows and columns are generally easier to analyze. For example, with such tables, if the larger frequencies of the table gather along the diagonal cells, this clearly indicates an association between the variables. Sometimes, however, the figures within a contingency table are quite difficult to interpret. This can happen for two main reasons: (1) the categories of one or both the variables are too numerous and/or uneven; (2) the frequencies and/or the percentages have no discernible pattern, because, for instance, the relationship between the variables is not linear. In the first case, it could be useful to aggregate or dichotomize the categories (this often happens in the case of Likert scale variables). In most cases, this solution leads to more readily interpretable results, though some information is lost in the process. In general, it is quite helpful to calculate the chi-square test or other measures of significance and/or association that summarize in a single figure the relationship between the variables.

Alberto Trobia

See also Chi-Square; Dependent Variable; Independent Variable; Likert Scale; Marginals; Nominal Measure; Ordinal Measure

Further Readings

De Vaus, D. A. (1996). *Surveys in social research.* London: UCL Press.

Gilbert, N. (1993). *Analyzing tabular data: Loglinear and logistic models for social researchers.* London: UCL Press.

CONTINGENT INCENTIVES

Past research has shown that contingent incentives can be used in survey research as a way of increasing survey response rates. The concept of contingent versus noncontingent incentives is that a noncontingent incentive is given to the respondent regardless of whether the survey task is completed, whereas giving a contingent incentive is dependent on the respondent's completion of the survey task, such as completing and returning the questionnaire in a mail survey. Contingent incentives are most commonly used with phone and Internet surveys, although they can be used with any mode of survey data collection. Usually the researcher will use the promise of the incentive as an inducement to coax the respondent into completing the survey, because the respondent does not receive the contingent incentive unless the survey task is completed.

The most common type of contingent incentive in survey research is the monetary incentive, most often paid either in the form of cash or in the form of a check. The recent introduction of cash cards and gift cards have made this form of monetary incentive another viable option for use in surveys. Some examples of nonmonetary contingent incentives include sweepstakes entries, charitable donations, videos, gas cards, coupons, online credits, small household appliances, books, electronic devices, small gadgets or knickknacks, and so on. However, research indicates that monetary contingent incentives are more effective than nonmonetary incentives of the same value.

Contingent incentives have generally been found to be less effective than noncontingent incentives for completing a survey. This often is the case even when the contingent (promised) incentive is several times larger in value than the noncontingent incentive given to a respondent before she or he completes the survey task. However, in some situations, it is impractical to offer a noncontingent incentive. Normally a noncontingent incentive would be offered in a situation in which there is an easy way to deliver it at the same time as the survey instrument, such as in a mailed survey. In contrast, the contingent incentive is, by definition, given after the survey task is completed. How soon after this is promised to take place will also affect the power of the contingent incentive to raise the response rate. The sooner the contingent incentive is given to the respondent after she or he completes the survey task, the greater its power to raise response rates. With telephone and in-person interviews, a contingent incentive can be a strong persuader for the interviewer to use to gain cooperation. However, in the case of a telephone survey, gratification in receiving the contingent incentive is delayed, unlike an in-person interview in which the incentive can be given immediately after the survey task is completed. Similarly, a monetary contingent incentive paid in cash provides more immediate gratification than one paid via check or cash card. Thus, contingent incentives paid in cash immediately upon completion of the survey task are likely to have the greatest positive impact on raising responses rates compared to contingent incentives of the same value that are given after some lag in time and/or are not given as cash.

The decision to use a contingent incentive is somewhat independent from the decision to use a noncontingent incentive. If the survey budget can afford both, researchers will still be somewhat at a loss as to how to distribute the total value that will be used across the noncontingent and contingent incentives. That is, there is no definite guidance provided by the research literature indicating what is the most optimal balance between the value of a contingent incentive and a noncontingent incentive when both are used in the same survey.

When considering which type of contingent incentive, if any, to use in a particular survey, the researcher should consider the type of survey instrument (mailed, phone, Internet, in-person), the relative importance of the response rate, the level of effort required to complete the survey, the probable motivation of the sample to comply without any incentive, and the need possibly to differentially incent certain hard-to-reach demographic cohorts. For simple, short mailed surveys, short phone interviews, and short Internet surveys, an incentive may not be needed. As the length and complexity of the survey increases or respondent engagement (e.g., level of interest) decreases, the need to consider the use of a noncontingent incentive is likely to increase.

The amount of contingent incentive offered to the respondent should not be out of proportion to the effort required to complete the survey. When a promised contingent incentive amount is the sole motivating factor in the decision of a respondent to cooperate, the respondent may put in a less than adequate effort in accurately and completely answering the questions in the survey. Researchers should be aware of this "buying cooperation" phenomenon. Some research organizations offer points for completing surveys that can later be redeemed for prizes. Some firms form panels of households that will complete numerous surveys and can accumulate points over time and redeem them for larger prizes.

Another use for contingent incentives is to persuade the participants to return all materials and do so in a timely manner. The participant may be motivated to make a deadline for returns if they are aware that the amount of the contingent incentive is at least partially dependent on returning the materials by the cutoff date.

A concern to some researchers who are considering use of a contingent versus noncontingent incentive with a mail or Internet survey (ones not administered by an interviewer) is the possibility of confusion about whether the survey task (e.g., questionnaire) was fully completed and returned in a timely manner.

Respondents may think that they did everything required to qualify for the incentive, while the researcher's records indicate otherwise. This confusion could cause both a public relations problem and a logistical nightmare for the survey organization if not properly handled. Thus researchers must ensure that clear and complete procedures and guidelines as well as contingency plans are established when using a contingent incentive. Any contingent incentive offer should be structured in such a way that the respondent is aware of what needs to be done to qualify for the incentive and that the researcher has a means of delivering that incentive in a reliable and straightforward way.

Norm Trussell

See also Economic Exchange Theory; Incentives; Leverage-Saliency Theory; Noncontingent Incentives; Social Exchange Theory

Further Readings

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: Wiley.

James, J. M., & Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality on mail surveys. *Public Opinion Quarterly*, *54*, 346–361.

Singer, E., Van Hoewyk, J., Gebler, N., Trivellore, T., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, *15*, 231–250.

Singer, E., Van Hoewyk, J., & Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, *64*, 189–205.

Control Group

In experimental designs, a control group is the "untreated" group with which an experimental group (or treatment group) is contrasted. It consists of units of study that did not receive the treatment whose effect is under investigation. For many quasi-experimental studies, treatments are not administered to participants, as in true experimental studies. Rather, treatments are broadly construed to be the presence of certain characteristics of participants, such as female gender, adolescence, and low socioeconomic status (SES), or features of their settings, such as private schools or participation in a program of interest. Thus, the control group in quasi-experimental studies is defined to be those lacking these characteristics (e.g., males, respondents who are older or younger than adolescence, those of high and medium SES) or absent from selected settings (e.g., those in public schools, nonparticipants in a program of interest). Control groups may alternatively be called "baseline groups."

In a true experiment, control groups are formed through random assignment of respondents, as in between-subject designs, or from the respondents themselves, as in within-subject designs. Random assignment supports the assumption that the control group and the experimental group are similar enough (i.e., equivalent) in relevant ways so as to be genuinely comparable. In true experimental studies and between-subject designs, respondents are first randomly selected from the sampling frame; then they are randomly assigned into either a control group or an experimental group or groups. At the conclusion of the study, outcome measures (such as responses on one or more dependent variables, or distributions on survey items) are compared between those in the control group and those in the experimental group(s). The effect of a treatment (e.g., a different incentive level administered to each group) is assessed on the basis of the difference (or differences) observed between the control group and one or more experimental group.

Similarly, in within-subject designs, respondents are randomly selected from the sampling frame. However, in such cases, they are not randomly assigned into control versus experimental groups. Instead, baseline data are gathered from the respondents themselves. These data are treated as "control data" to be compared with outcome measures that are hypothesized to be the result of a treatment after the respondents are exposed to the experimental treatment. Thus, the respondents act as their own control group in within-subject designs.

Control groups are often used in evaluation studies that use surveys, and they are also relevant to methodological research on surveys. Research that examines the effects of questionnaire design, item wording, or of other aspects of data collection often uses a classical "split-ballot" design or some variant. In these studies, respondents are assigned at random to receive one of two versions of a questionnaire, each version varying on a single point of question order, wording, or presentation. In practice, these studies often depart from the conception of presence versus absence that typically marks the contrast between treatment and control groups. Researchers may present a variation of an item to both groups, for example, as opposed to administering the item to one group and not to the other. Nevertheless, these lines of survey research rely on the control group—either literally or by extension—as a necessary support for claims about the causal effects of the items, procedures, or programs being studied.

Chao-Ying Joanne Peng and Mary B. Ziskin

See also Experimental Design; Factorial Design; Random Assignment; Split-Half

Further Readings

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Newbury Park, CA: Sage.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Huck, S. (2004). *Reading statistics and research* (4th ed.). New York: Pearson Education.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.

Marsh, C. (1982). *The survey method: The contribution of surveys to sociological explanation*. London: George Allen & Unwin.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology*, *34*, 311–347.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology*, *22*, 377–399.

Controlled Access

Any sampled housing unit to which access by a data collector is physically blocked or impeded is considered to be a situation of controlled access. Impediments may include people (e.g., a "gatekeeper"), structures, and/or animals. Controlled access situations are encountered only in studies using the in-person field data collection methodology. Dealing effectively with these impediments is necessary to further the objectives of a field data collection operation.

Controlled access situations can take many forms and may involve one impediment or multiple impediments occurring simultaneously. For example, a single-family home may be surrounded by a locked fence or may have a growling dog loose in the yard, or both. A secured apartment building may have a locked entrance, a security guard, or both. An entire residential neighborhood may have keycard access–only gated entrances.

It is important to consider that controlled access situations may involve not just one but multiple sampled housing units. For example, in the case of an area probability sample, a locked apartment building may encompass a number of sampled units.

Security features that impede access to housing units are not limited to particular socioeconomic areas. High-crime, lower–socioeconomic status areas may have more gated yards with guard dogs, bars on windows and doors, and locked apartment buildings. More affluent areas may have gates on the street and/or driveway entrances, security guards, and locked apartment buildings.

Another example of controlled access situations affecting multiple sample units is group quarters. University dormitories, military barracks, and other institutionalized living units are primary examples. Similarly, in the United States, Native American Indian reservations often present controlled access challenges.

Addressing controlled access situations will generally fall into one of two approaches: *overt* or *covert*. Covert methods often are more efficient and effective provided they do not put the data collector in legal or physical jeopardy. One example would be following a resident into a locked apartment building when he or she open the door. Another would be, once a selected unit resident grants entrance to the building over the intercom, using that access to go to all other selected units in the building. Overt methods, however, may be the only practical means of dealing with some situations. This may involve sending letters and/or making presentations to the controllers (gatekeepers) of the physical barrier (e.g., building manager, homeowners' or tenants' association).

Regardless of the type of intervention, success will depend first on gathering sufficient, detailed information about the situation. After analyzing the information, appropriate options and strategies must be devised and implemented. Although it is sometimes better to "beg forgiveness later than ask permission first," it may be advisable to require field data collectors to consult with their supervisors before using a covert method of gaining entry to a controlled access environment.

Researchers should include in their procedural manuals and training programs material on how to deal effectively with various controlled access situations. Strategies and tools for dealing with locked facilities, complexes, and neighborhoods should be developed, utilized, and continually enhanced in an effort to negotiate past these impediments. This is particularly important so that data collectors do not find themselves taking unnecessary risks. They must be prepared to exercise good judgment to avoid legal issues such as trespassing or being injured attempting to surmount a physical barrier or outrun an aggressive animal.

As our society becomes increasingly security- and privacy-minded, the presence of controlled access situations and facilities will similarly increase. It is important for researchers to recognize this trend and the potential negative effect on survey nonresponse that controlled access situations represent.

Randall Keesling

See also Face-to-Face Interviewing; Field Survey; Field Work; Gatekeeper

CONTROL SHEET

A control sheet, also called a "case control form," is used by interviewers in in-person (face-to-face) surveys to record information about the contact attempts they make with households or persons who have been sampled.

Similar in purpose to the call sheet used by telephone interviewers, the control sheet captures key paradata about each contact attempt an interviewer makes with the household or person. This includes (a) the date of the contact attempt, (b) the time of day of the contact attempt, (c) the outcome (disposition) of the contact attempt, and (d) any additional information that is pertinent about the effort to make contact (e.g., the name of the designated respondent if she or he is not home at the time the attempt is made and the best time to recontact her or him).

The information recorded on control sheets serves several important purposes. First, it allows the interviewers and supervisory field staff to better control the processing of the sample according to the a priori contact rules that have been established by the researchers. For example, these rules set guidelines about how many times a person or household can be contacted within a week's period; how many of these contacts should be during the day on weekdays, in the evening hours of weekdays, or on weekends; and how many days must elapse between a first refusal and an attempt to convert the refusal. The control sheet is the mechanism that brings order to the systematic processing of the sample. Second, the information on the control sheet about previous contact attempts allows an interviewer to be better prepared to gain a completed interview the next time she or he tries to contact the household. Third, the information on the control sheet can be used by supervisory staff in their ongoing and annual evaluations of the performance of individual interviewers, teams of interviewers, and/or the interviewing staff as a whole. Fourth, the information on the control sheet can be analyzed by the researchers to investigate ways to improve the cost-effectiveness of future interviewing (e.g., studying the optimal time lapse between a first refusal and a successful conversion attempt).

Paul J. Lavrakas

See also Calling Rules; Call Sheet; Dispositions; Face-to-Face Interviewing; Field Survey; Field Work; Paradata; Refusal Conversion; Refusal Report Form (RRF); Standard Definitions; Supervisor

CONVENIENCE SAMPLING

Convenience sampling is a type of nonprobability sampling in which people are sampled simply because they are "convenient" sources of data for researchers. In probability sampling, each element in the population has a known nonzero chance of being selected through the use of a random selection procedure. Nonprobability sampling does not involve known nonzero probabilities of selection. Rather, subjective methods are used to decide which elements should be included in the sample. In nonprobability sampling, the population may not be well defined. Nonprobability sampling is often divided into three categories: purposive sampling, convenience sampling, and quota sampling.

Convenience sampling differs from purposive sampling in that expert judgment is not used to select a representative sample of elements. Rather, the primary selection criterion relates to the ease of obtaining a sample. Ease of obtaining the sample relates to the cost of locating elements of the population, the geographic distribution of the sample, and obtaining the interview data from the selected elements. Examples of convenience samples include mall intercept interviewing, unsystematically recruiting individuals to participate in the study (e.g., what is done for many psychology studies that use readily available undergraduates), visiting a sample of business establishments that are close to the data collection organization, seeking the participation of individuals visiting a Web site to participate in a survey, and including a brief questionnaire in a coupon mailing. In convenience sampling the representativeness of the sample is generally less of a concern than in purposive sampling.

For example, in the case of a mall intercept survey using a convenience sample, a researcher may want data collected quickly using a low-cost method that does not involve scientific sampling. The researcher sends out several data collection staff members to interview people at a busy mall, possibly on a single day or even across a weekend. The interviewers may, for example, carry a clipboard with a questionnaire that they may administer to people they stop in the mall or give to people to have them fill out. This variation in convenience sampling does not allow the researcher (or the client) to have any sense of what target population is represented by the sample. Although convenience samples are not scientific samples, they do on occasion have value to researchers and clients who recognize their severe limitation; for example, they may allow some quick exploration of a hypothesis that the researcher may eventually plan to test using some form of probability sampling.

Mike Battaglia

See also Mall Intercept Survey; Nonprobability Sampling; Probability Sample; Purposive Sample

Further Readings

Henry, G. (1990). *Practical sampling.* Newbury Park, CA: Sage.

Convention Bounce

Support for presidential candidates usually spikes during their nominating conventions—a phenomenon so reliable its measurement has become a staple of pre-election polling and commentary. Some of these *convention bounces* have been very short-lived, the race quickly reverting to its pre-convention level between the candidates. Others have been more profound—a coalescing of voter preferences that has charted the course for the remaining campaign.

While convention bounces have been apparent since 1968 (previous election polling was too infrequent for reliable identification of such bounces), focus on the convention bounce owes much to Bill Clinton, who soared from a dead heat against Republican presidential incumbent George H. W. Bush before the 1992 Democratic convention to nearly a 30-point lead after it. While the race later tightened, Clinton never again trailed in pre-election polls.

No bounce has matched Clinton's, but others are impressive in their own right. Jimmy Carter rode a 16-point bounce to a 33-point lead after the 1976 Democratic convention, lending authority to his challenge and underscoring incumbent Gerald Ford's weakness. Ford in turn mustered just a 7-point bump following the 1976 Republican convention; while the race tightened at the close, Carter's higher bounce foretold his ultimate victory.

If a solid and durable bounce suggests a candidate's strength, its absence can indicate the opposite. Neither Hubert Humphrey nor George McGovern took significant bounces out of their nominating conventions in 1968 and 1972, both en route to their losses to Richard Nixon.

Assessment

Standards for assessing the bounce differ. While it sometimes is reported among "likely voters," it is more meaningfully assessed among all registered voters, which is a more stable and more uniformly defined population. And the fullest picture can be drawn not by looking only at change in support for the new nominee, but—offense sometimes being the best defense in politics—at the change in the margin between the candidates, to include any drop in support for the opposing candidate. For example, the 1968 Republican convention did more to reduce Humphrey's support than to bolster Nixon's.

Timing can matter as well; surveys conducted closer to the beginning and end of each convention better isolate the effect. In 2004, Gallup polls figured John Kerry's bounce from a starting point measured 5 days before his convention began and assigned him a net loss of 5 points—its first negative bounce since McGovern's 32 years earlier. Using different timing, ABC News and *The Washington Post* started with a pre-convention measurement done 4 days later than Gallup's, and found an 8-point bounce in Kerry's favor, much nearer the norm.

Using the change in the margin, among registered voters, the average bounce has been 10 points in Gallup polls from 1968 through 2004 (and, for comparison, a similarly sized bounce of 13 points in ABC News polls from 1992 to 2004). While individual bounces vary, on average they have been consistent across a range of parameters: in Gallup data, 11 points for Democratic candidates (9 points leaving aside Clinton's 1992 bounce), 9 points for Republicans, 8 points for incumbents, 11 points for challengers, 10 points for better-known candidates (incumbent presidents and incumbent or former vice presidents), 10 points for less-known candidates, 12 points after each cycle's first convention, and 9 points after the second convention.

While the average size of the bounces by the candidate's political party are similar, more of the drama has been among Democratic candidates—a standard deviation of 10 in their bounces (8 without Clinton's in 1992) compared with 4 in the Republicans'. The average Democratic bounce correlates significantly with the average bounce overall, while the average Republican bounce does not.

Causes

The basis for the bounce seems clear: a specific candidate dominates political center stage for a week, laying out his or her vision, burnishing his or her credentials and—directly or through surrogates—criticizing his or her opponent. It takes a problematic candidate, an off-key convention, or an unusually immovable electorate not to turn the spotlight into support.

But exposure is not the sole cause; while airtime for network coverage of the conventions has declined sharply over the years, the bounces haven't. The two national conventions received a total of 73 hours of broadcast network coverage in 1968, declining sharply in ensuing years to a low of 6 hours in 2004 (as reported by Harold Stanley and Richard Niemi in *Vital Statistics on American Politics* 2003–2004). Audience ratings likewise dropped. Yet there is no significant relationship between hours of network coverage and size of convention bounces. Indeed, the largest bounce on record, Bill Clinton's in 1992, occurred in the modern era of less network news coverage—8 hours for his convention—while George McGovern's bounceless 1972 convention was one of the most heavily covered, at 37 hours.

A range of other factors may contribute to the bounce. Vice presidential running mates often are named during or shortly before conventions. Events outside the convention doors can play a role, such as the Chicago riots of 1968 or the on-again, off-again Ross Perot candidacy of 1992 (although data from that time indicate that Perot was more a casualty of Clinton's convention surge than its cause). Strength of support is another factor, informed by the level of political polarization or the extent of economic discontent heading into the convention season. And atop the heap stands the effectiveness of the individual candidates and their campaigns.

As to why there is more variability in Democratic bounces, causal influences may include the objective quality of individual candidates, a generally declining Democratic advantage in partisan self-identification across this period, and perhaps, more steadfast support among Republican self-identifiers for their party's nominees.

Whatever the other influences, presidential nominating conventions mark unique and highly fraught periods in the election cycle, when public attention focuses, candidates pass—or fail to clear—the basic bar of acceptability to a broader audience, and their support often undergoes its biggest swings of the contest.

The varying size of convention bounces suggests that they are founded on evaluative assessments, not simply the quantity of news coverage. The fact that some bounces fade rapidly while others endure similarly underscores the substance of what is occurring beneath the bright lights and balloons. A focusing of the public's attention may inspire the bounce, but a more deliberative judgment determines its size, staying power, and ultimate impact on Election Day.

Gary Langer

See also Election Polls; Horse Race Journalism; Likely Voter; Media Polls; Pre-Election Polls

Further Readings

Stanley, H. W., & Niemi, R. (2003). *Vital statistics on American politics 2003–2004.* Washington, DC: Congressional Quarterly Press.

Conversational Interviewing

Conversational interviewing is also known as "flexible" interviewing or "conversationally flexible" interviewing. These terms refer to an alternative style of survey interviewing that allows deviations from the norms of standardized interviewing. Under conversational interviewing procedures, interviewers are allowed to ask respondents if they did not understand a question and provide unscripted feedback to clarify the meaning of questions as necessary. Conversational interviewing represents an alternative set of techniques to standardized survey interviewing whereby interviewers are allowed to provide unscripted information to respondents in an effort to clarify question meaning.

Proponents of conversational interviewing techniques argue that standardized procedures may reduce the accuracy of survey responses because standardization precludes conversational interactions that may be required for respondents to understand some questions. A key distinction between standardized and conversational interviewing is that standardization requires the interpretation of questions to be accomplished entirely by respondents. A central tenet of standardized interviewing is that interviewers must always read questions, response options, and instructions to respondents exactly as they are scripted. Further definitions, clarifications, or probes can only be read in standardized interviews if these elements are included in the interview script. A second tenet of standardized interviewing is that any probes used by interviewers must be nondirective, so that the probes do not lead respondents to give particular answers. As a result, standardized interviewers can only provide clarification when respondents request it, and can then only provide standardized forms of assistance such as nondirective probes.

In conversational interviewing, interviewers can provide whatever information is needed to clarify question meaning for respondents, and they can provide these clarifying statements whenever they perceive respondents are having difficulty understanding a question. Proponents of conversational interviewing hypothesize that these more flexible techniques can produce more accurate survey responses by standardizing the *meaning* of questions, not the wording or exact procedures used to administer the questions. Because the same terms can have different meanings to different respondents, conversational interviewing may improve response accuracy by allowing unscripted exchanges between interviewers and respondents to clarify the meaning of specific terms. Based on this reasoning, conversational interviewing techniques are assumed to increase the accuracy of survey responses, particularly in those situations in which respondents cannot initially map the specific terms in a question to the relevant information they have to report.

Experimental studies have been conducted to assess whether more flexible conversational interviewing techniques could produce more accurate data than standardized procedures for some survey questions. In these experiments, respondent interviews were assigned either to a standardized condition in which interviewers were not allowed to deviate from the script or to a conversational condition in which interviewers were allowed to encourage respondents to ask questions if they did not understand and provide unscripted feedback to clarify the meaning of question terms.

Results of this research indicated that the two alternative interviewing procedures both produced nearly perfect accuracy when question concepts clearly mapped onto the situations respondents had to report. For example, respondents were asked about purchasing furniture, so those who had purchased items like tables and chairs could clearly map their situation onto the question concept and accurately answer this question with either interviewing procedure. In contrast, respondents who had purchased an item such as a lamp, for example, could not clearly answer the question about purchasing furniture. In interviews in which question concepts did not clearly match respondents' situations, conversational interviewing procedures increased response accuracy by nearly 60%. Additional research indicated that data from follow-up interviews using conversational techniques increased the accuracy of reports compared to an initial round of standardized interviews. In addition, respondents in this experiment were twice as likely to change their answers between a first standardized interview and a second conversational interview (22%) than between a first standardized interview and a second standardized interview (11%). The results of these experiments generally confirmed that conversational techniques led to greater response accuracy when ambiguity existed between the key concepts of the question and the information respondents had to report.

Successfully applying conversational interviewing techniques in social surveys remains limited by a few important considerations. First, research has not yet demonstrated whether large numbers of interviewers can be trained and supervised effectively to apply conversational techniques in a way that does not introduce other kinds of response bias. Research to date has involved only a small number of interviewers and a limited number of interviews in which interviewer training and procedures could be tightly controlled. A second limitation is that research has indicated conversational interviews improves response accuracy compared to standardized interviews only for questions in which considerable ambiguity exists. Most of the questions developed, tested, and implemented in various surveys are not subject to the same degree of ambiguity required to produce benefits from conversational techniques. Third, using conversational interviewing procedures increased the average interview length in experimental studies by 80% compared to administering the same set of questions with standardized techniques. Conversational interviewing may produce more accurate data than standardized interviewing for some survey items, but the more flexible interviewing conditions limit the number of survey items that can be asked in the same interview time.

Douglas B. Currivan

See also Cognitive Aspects of Survey Methodology (CASM); Interviewer Effects; Interviewer-Related Error; Interviewer Training; Interviewer Variance; Interviewing; Nondirective Probing; Probing; Standardized Survey Interviewing

Further Readings

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, *64*, 1–28.

Maynard, D. W., Houtkoop-Steenstra, H., Schaeffer, N. C., & Van der Zouwen, J. (Eds.). (2002). *Standardization and tacit knowledge: Interaction and practice in the survey interview.* New York: Wiley.

Schaeffer, N. C. (1991). Conversation with a purpose—or conversation? Interaction in the standardized interview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367–391). New York: Wiley.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, *61*, 576–602.

COOPERATION

Cooperation is a term used by survey researchers that refers to the degree to which persons selected (sampled) to participate in research accept (agree to) their invitation and engage (cooperate) in the research process. The composition of the group under study is a fundamental (and vitally important) consideration in the design, execution, and interpretation of a survey. A researcher must both identify and collect information from an appropriate sample in order to successfully and validly answer the research question. Ideally, the rate of cooperation among those sampled will be very high.

Applied to a specific study, cooperation refers to the breadth of participation that researchers are able to elicit from those that they have chosen to study. To help objectively measure levels of cooperation within a study, the American Association for Public Opinion Research (AAPOR) developed a series of standard definitions that include how to define and compute cooperation rates. AAPOR's cooperation rates are mathematical formulae that reflect the proportion of respondents who actually participate in a survey divided by all of the sampled cases that are ever contacted, and are eligible, to participate in the survey.

Together with the response, refusal, and contact rates, the cooperation rate is included in a category of formulas collectively known as "outcome rates." These rates are calculated by survey researchers in order to better understand the performance of surveys. Methods sections of survey reports typically include at least some information regarding these rates.

Factors Affecting Cooperation

There is a wide body of literature regarding the theory, application, and relationship of the factors that affect cooperation. Examples of the major types of factors that can affect cooperation include the following:

- Level of effort used in recruiting respondents
- Respondents' interest in the topic of the survey
- Study's mode of data collection
- Skill of interviewers in interviewer-administered surveys
- Information given to respondent prior to his or her engaging in survey
- Length/burden of the survey

- Whether or not incentives are offered
- Characteristics of the population of interest

Cooperation in Random Samples

Statistical theory explains that data should be collected from all those selected for inclusion (sampled) in probabilistic samples. In practice, this is seldom achieved. Any individual who is selected but does not participate in a study is termed a "nonrespondent" and may (or may not) induce nonresponse bias. One possible scenario, for example, is that the data from a survey yielding poor cooperation levels may be heavily distorted if nonresponders differ systematically in nonnegligible ways from responders.

Although there is common agreement that general cooperation levels within the United States have been in a state of decline for years, many within the survey research community believe that poor cooperation levels have been overstated as a threat to validity in random samples. Nevertheless, cooperation continues to be viewed as one of the important indicators of the performance of a survey and is properly considered in the context of both the study's target population and variables of interest.

The term *cooperation* is strongly associated with probabilistic samples in quantitative surveys because of its connection to the validity of random samples. However, cooperation plays an important role in both quantitative and qualitative research.

Society and Cooperation

In its broadest sense, cooperation is often discussed in the context of the overall state, or health, of survey research. From this perspective, survey research professionals are concerned with how society perceives survey research as an activity or "enterprise." For example, an atmosphere of low cooperation in society may reflect dissatisfaction with research (or research techniques) among the public, which in turn, may result in legislation that restricts or inhibits survey and opinion research.

CMOR, the Council for Marketing and Opinion Research, operates to promote respondent cooperation and protect and promote government affairs on behalf of the survey research profession. CMOR stresses that a critical step in improving general respondent cooperation includes researchers universally adopting practices that foster a favorable relationship between research and the public. To this end, CMOR has published and encourages all researchers to adhere to the Respondent Bill of Rights. It also encourages members of the profession to use the same outcome rate calculations to ensure that there are consistent measures in the profession.

Patrick Glaser

See also American Association for Public Opinion Research (AAPOR); Cooperation Rate; Council for Marketing and Opinion Research (CMOR); Incentives; Leverage-Saliency Theory; Nonresponse; Nonresponse Error; Respondent Burden; Standard Definitions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Council for Marketing and Opinion Research. (2008). *Respondent bill of rights*. Glastonbury, CT: Author. Retrieved March 24, 2008, from http://www.cmor.org/research/rights.cfm

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly*, *64*, 299–308.

Cooperation Rate

The cooperation rate to a survey indicates the extent to which contacted individuals cooperate with a request to participate in a survey. It is often mistakenly reported or interpreted as the response rate. Generally, the cooperation rate is the ratio of all cases interviewed out of all eligible units ever contacted, whereas a response rate is the ratio of all cases interviewed out of all eligible sample units in the study, not just those contacted.

The American Association for Public Opinion Research (AAPOR), which has established a standard definition of the cooperation rate, offers at least four ways to calculate it. The numerator includes all completed interviews but may or may not include partial interviews. The denominator includes all eligible sample units that were contacted (including refusals and other non-interviews that may have been contacted), but may or may not include sample units that are incapable of cooperating (e.g., because of health or language barriers).

When reporting the cooperation rate, researchers should clearly define the rules for survey eligibility and explain how they decided to calculate the rate. The level at which the rate has been calculated (individual, household, school district, business, etc.) should be reported. Though cooperation rates are most often calculated using only contacts with known eligible respondents, if there is a screener, consumers of survey results might also want to know the percentage of people who cooperate with the screener in addition to the percentage of people who participated in the full survey.

One important variation in how the cooperation rate is calculated is whether contacted sample members with unknown eligibility are included in the denominator of the calculation. It is possible to include in the denominator an estimate of all eligible cases (or *e*, the proportion of cases with unknown eligibility assumed to be eligible), not just the cases confirmed as eligible.

A lower cooperation rate implies a lower response rate, raising concerns about the representativeness of the participating sample members. For example, Robert Groves and Mick Couper report that some research has shown that noncooperating sample members score lower on social engagement indices than do cooperating sample members. If measures of social engagement are important analytical variables, then a low cooperation rate may bias survey estimates.

The cooperation rate also has implications for survey costs, as it is an indicator of sample yield (i.e., the number of completed interviews achieved from a fixed number of sample units). The lower the cooperation rate, the more the effort needed to achieve a required number of completed interviews, whether that effort involves enlarging the sample, making additional contacts to sample members, training interviewers, or providing incentives to increase cooperation. For interviewer-administered surveys, the cooperation rate serves as one measure of the interviewer's success.

Survey organizations try to maximize the response rate by maximizing the cooperation rate (in addition to maximizing the contact rate, or the proportion of all sample members for which a person was reached). For instance, researchers may try to alter the sample members' predisposition toward survey participation by changing the nature of the initial contact to make the survey more appealing. Very often, cooperation is manipulated through advance mailings and through the interviewer. The issue of interviewer–respondent interaction and its influence on survey cooperation has received considerable attention in the recent literature on survey research, thus motivating survey organizations to focus on interviewer training. The training generally emphasizes avoiding refusals, tailoring the interview approach to sample members, and maintaining the interaction with sample members while on the telephone or at the doorstep. Evidence from studies of interviewer training and interviewer–respondent interactions suggests that tailoring and maintaining interaction are important to maximizing cooperation rates.

Danna Basson

See also American Association for Public Opinion Research (AAPOR); Contact Rate; Cooperation; *e*; Interviewer Training; Leverage-Saliency Theory; Refusal Rate; Response Rates; Standard Definitions; Survey Costs; Tailoring; Unknown Eligibility

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2001). *Survey nonresponse.* New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

CORRELATION

Correlation is a statistical measure of the relationship, or association, between two or more variables. There are many different types of correlations, each of which measures particular statistical relationships among and between quantitative variables. Examples of different types of correlations include Pearson's correlation (sometimes called "product-moment correlation"), Spearman's correlation, Kendall's correlation, intraclass correlation, point-biserial correlation and others. The nature of the data (e.g., continuous versus dichotomous), the kind of information desired, and other factors can help determine the type of correlation measure that is most appropriate for a particular analysis.

The value of the correlation between any two variables is typically given by a correlation coefficient, which can take on any value between and including −1.00 (indicating a perfect negative relationship) up to and including +1.00 (indicating a perfect positive relationship). A positive correlation between two variables means that as the value of one variable increases, the value of the second variable tends to increase. A negative correlation means that as the value of one variable increases, the value of the second variable tends to decrease. A correlation that is equal to zero means that as one variable increases or decreases, the other does not exhibit a tendency to change at all.

One frequently used measure of correlation is Pearson's correlation; it measures the linearity of the relationship between two variables. The Pearson's correlation coefficient is calculated by dividing the covariance of two variables by the product of the standard deviation of each variable. That is, for n pairs of variables x and y, the value of the Pearson's correlation is

$$\frac{\sum_{i=1}^{n}[(x_i-\bar{x})\times(y_i-\bar{y})]}{n}\Bigg/\left(\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}\times\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}\right).$$

For instance, as part of a study on smokers' health and demographics, a survey researcher might collect data on smokers' annual household income and the average number of cigarettes smoked daily. The data for 10 smokers—sorted in ascending order of income—might look like Table 1.

In this case, simple inspection reveals that the correlation is negative. That is, as income increases, the average number of cigarettes smoked daily tends to decrease. The value of the Pearson's correlation between these variables equals –0.484, confirming that the relationship between the two variables is, in fact, negative and moderately linear. A scatter plot of these variables visually illustrates the nature of this relationship, as shown in Figure 1 (next page).

While correlation analysis describes one aspect of the quantitative relationship between variables, it certainly has its limitations. First, it cannot be used to infer the extent of a causal relationship. For example, the preceding example shows only that income and average number of cigarettes smoked daily for these 10 individuals are related in a negative, somewhat linear fashion.

Table 1 Cigarettes and income

Average Number of Cigarettes Smoked / Day	*Yearly Household Income (in $1,000s)*
10	28
30	31
5	45
25	46
22	48
12	55
4	57
13	62
4	62
8	77

It does not mean that increasing a smoker's income would cause a reduction in the number of cigarettes smoked or that smoking fewer cigarettes would cause an increase in an individual's income.

A second important limitation is that correlation analysis does not provide any information about the magnitude—or the size—of the relationship between variables. Two variables may be highly correlated, but the magnitude of the relationship might, in fact, be very small. For instance, the correlation of –0.484 between income and average number of cigarettes smoked daily in the example says only that the relationship is negative and that the relationship is somewhat linear. It does not provide any information regarding how many fewer cigarettes are related to an increase in income. That is, every extra dollar of income could be associated with a decrease in average number of cigarettes that is very large, very small, or anywhere in between.

Joel K. Shapiro

See also Noncausal Covariation; ρ (Rho); Standard Error; Variance

Further Readings

Kvanli, A. H., Guynes, C. S., & Pavur, R. J. (1986). *Introduction to business statistics.* St. Paul, MN: West.

Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics.* New York: Wiley.

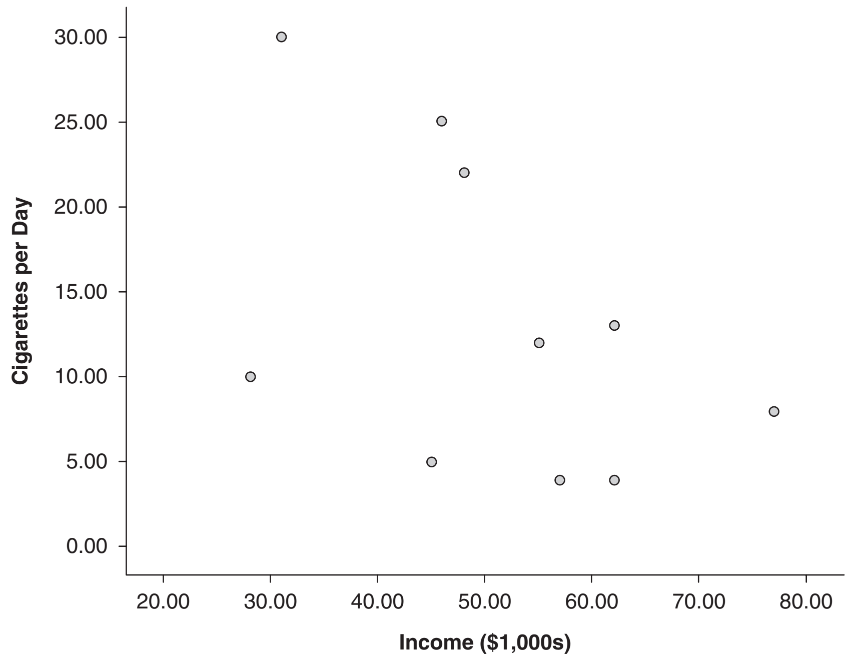

Figure 1 Cigarettes versus income

COUNCIL FOR MARKETING AND OPINION RESEARCH (CMOR)

The Council for Marketing and Opinion Research (CMOR) is a national nonprofit organization founded to work on behalf of the marketing and opinion research industry in two key areas:

1. To improve respondent cooperation across all modes of survey data collection and focus groups
2. To promote positive state and federal legislation that affects marketing and opinion research, to monitor and prevent restrictive legislation that has the potential to impact research work, and to encourage self-regulation among the survey research profession

CMOR was founded in 1992 by four of the major marketing research trade associations: AMA (American Marketing Association), ARF (Advertising Research Foundation), CASRO (Council of American Survey and Research Organizations), and MRA (Marketing Research Association). These organizations believed that the two areas of focus—respondent cooperation and government affairs—were so critical to the research industry that a specialized industry group should be created to devote attention and solutions to these research issues.

CMOR is composed of more than 150 organizations that represent all facets of the research profession:

- Client companies (or end users of research)
- Full-service research companies
- Data collection companies
- Other associations in the profession
- Academic institutions
- Government entities
- Research-related services (such as sampling and software companies)

Organizational Structure

A volunteer board of directors and volunteer committee set CMOR's policy and vision and determine the direction of CMOR's initiatives. Members are drawn from all sectors of the research industry: full-service research firms, data collection companies, research analysts, and end users.

CMOR is structurally organized into two separate departments: Respondent Cooperation and Government

Affairs. Each department maintains a permanent volunteer committee in order to drive the organization's work. Additional committees are formed on an ad hoc basis. A professional staff person oversees both departments and acts as a liaison with counterparts in the other research organizations. Further, professional staffers are hired to head each department and support staff assist in implementing the initiatives.

Respondent Cooperation and Government Affairs are inextricably related due to government's influence, through legislation, over what methods for conducting research are deemed legal and how this may affect the validity of research and the ability of the researcher to achieve respondent cooperation. Conversely, Respondent Cooperation is partially a reflection of the public's perceptions and attitudes toward research, and it may play a very strong role in the types of legislation that are proposed and adopted as law.

Respondent Cooperation

With regard to respondent cooperation, CMOR's mission is to evaluate the public's perceptions of the research process, to measure the effects of alternative methods of improving respondent cooperation, and to provide a foundation upon which to build an improved set of industry guidelines. Since its formation, CMOR has worked to increase respondent cooperation and has advocated the importance and necessity of marketing and opinion research to the general public. Objectives related to respondent cooperation objectives include the following:

- Provide objective information about level of cooperation in surveys
- Monitor the ever-changing research environment
- Develop industry-accepted and -supported solutions to improve respondent relations
- Educate and develop training programs for our members and members of the research community about the issues affecting respondent cooperation and of CMOR's efforts to improve participation
- Educate the research community's external audiences, including the public, media, and businesses, about the value of research and their participation in legitimate research surveys and polls
- Promote the social utility and value of survey research
- Act quickly to provide guidance to our members and the research community about environmental issues that may affect cooperation

Government Affairs

In terms of government affairs, CMOR's mission is to monitor relevant legislative and regulatory activity, to ensure that the interests of the research community are protected, and to educate industry members regarding relevant legislative, statutory, and legislative issues. The following are among the objectives in this area:

- Monitor and respond to legislative and regulatory activities that affect the research industry
- Educate CMOR members and members of the research community about the legislative and regulatory measures that threaten research and of CMOR's efforts to protect the research industry
- Educate CMOR members and members of the research community about existing statutes and regulations that impact the research industry
- Educate lawmakers and policymakers about the value of research, the distinction between research and sales-related activities and the negative implications restrictive measures have on research
- Respond to abuses of the research process and work with lawmakers and government officials to regulate and prosecute such activities
- Act pro-actively on legislative and regulatory measures
- Build coalitions with other organizations to use as resources of information and to strengthen our ability to act on restrictive and proactive legislative and regulatory measures

Kathy Pilhuj

See also Council of American Survey and Research Organizations (CASRO); Federal Communication Commission (FCC) Regulations

Further Readings

Council for Marketing and Opinion Research: http://www.cmor.org

Council of American Survey Research Organizations (CASRO)

The Council of American Survey Research Organizations (CASRO) is the national trade association for survey research businesses, whose 300-plus member companies (predominantly in the United States, but also in Canada, Mexico, and abroad) represent about 80% of the U.S. annual revenues in survey research businesses. Established in 1975, CASRO advances the business of research through standards, guidelines, professional development, and self-regulation in the process and performance of survey research. CASRO's mission is to provide the environment and leadership that will promote the profitable growth and best interests of those firms and other entities engaged in the survey research industry.

Standards and Guidelines

CASRO standards and guidelines provide mandatory and recommended processes and practices in survey research that ensure the quality and integrity of the survey research conducted by all CASRO members. CASRO's Code of Standards and Ethics for Survey Research, which is mandatory for all members, describes its members' responsibilities to respondents, to clients, and to the public. CASRO is the U.S. delegate (along with the American National Standards Institute) to the International Standards Organization's (ISO) planned development of a quality standard for market, opinion, and social research.

Professional Development

CASRO University is a professional development curriculum that provides certificates in Survey Research Practice, Business Management, Project Management, and Privacy Management. CASRO University includes an annual series of conferences, workshops, Webcasts, and other professional development and educational programs that contribute to the career development of survey researchers. CASRO and CASRO University work in cooperation with academic programs as well, including the graduate degree programs in survey research at the University of Georgia (Athens), University of Texas (Arlington), University of Wisconsin (Madison), Southern Illinois University (Edwardsville), and the Market Research Institute International (MRII). CASRO Financial Reports include annual Financial and Compensation Surveys, as well as an annual Data Collection Survey.

Self-Regulation

The CASRO Government and Public Affairs (GPA) program monitors, lobbies as appropriate, and provides guidance on compliance with legislation and regulations that impact survey research. In addition, the CASRO GPA proactively protects professional survey research from abuses and misuses such as "SUGing" (selling under the guise of research) and "FRUGing" (fundraising under the guise of research). The mission of CASRO GPA is to promote continued self-regulation, to encourage and support professional accountability, and to foster and ensure public trust.

Diane Bowers

See also American Association for Public Opinion Research (AAPOR); Council for Marketing and Opinion Research (CMOR); FRUGing; SUGing

Further Readings

Council of American Survey Research Organizations: http://www.casro.org

COVARIANCE

Covariance is a measure of association between two random variables. It has several applications in the design and analysis of surveys.

The covariance of two random variables, X and Y, is equal to the expected product of the deviations between the random variables and their means:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

Under a design-based perspective to surveys, the sample inclusion indicators are random variables, and covariance is present when the probabilities of inclusion are correlated.

For a simple random sample of n units from a population of size N, the covariance between the means $\bar{x}$ and $\bar{y}$ is estimated as:

$$cov(\bar{x},\bar{y}) = \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

This is equivalent to the variance formula when x_i and y_i are the same for each unit in the sample. For complex sample surveys, standard variance estimation techniques, such as Taylor series linearization, balanced repeated replication, or jackknife replication, can be used to compute covariance.

Covariance can be written as a function of the correlation $\rho(x,y)$:

$$cov(x,y) = \rho(x,y)var(x)var(y),$$

where var(x) and var(y) are the variances of x and y, respectively. The covariance of x and y is equal to zero when x and y are uncorrelated, as is the case when they are derived from two independent samples or from independent strata within the same sample. However, in many situations in sample surveys, the covariance is present and should not be ignored.

For example, suppose a nonresponse bias analysis is conducted to determine the impact of a low response rate on survey estimates. The bias in an estimate is

$$bias(y_R) = y_R - y,$$

where y_R is the estimate based on only the respondents and y is the estimate from the entire sample. The variance of the bias is

$$var(bias(y_R)) = var(y_R) + var(y) - 2 * cov(y_R, y).$$

In general, the variance of a linear combination of random variables, X_1 through X_n, is

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_i \sum_j a_i a_j Cov(X_i, X_j).$$

The percentage of females in the population is estimated as 48% based on only respondents but as 50% from the full sample, for a bias of -2%. Using the appropriate variance estimation method, the variances are found to be 1.2 for the estimate from respondents and 1.0 for the full sample, with a covariance of 0.9. Taking into consideration the correlation between estimates from the full sample and estimates from respondents only, the variance of the bias is 0.4 $(= 1.2 + 1.0 - (2^*0.9))$. Using a t-test to test the null hypothesis that the bias is equal to zero, the p-value is found to be < 0.001, indicating significant bias in the estimate of females. However, if the covariance term is ignored, the variance of the bias is calculated as 2.2, and the bias is no longer determined to be statistically significant.

Ignoring the covariance term leads to an overestimation of the variance of the difference of the estimates, given the two estimates are positively correlated. This result is important in other survey contexts, such as comparing estimates between two time periods for a longitudinal survey or from different subdomains involving clustering. Covariance also has several other applications in surveys, including intraclass correlations, goodness-of-fit tests in a regression analysis, and interviewer effects.

Wendy Van de Kerckhove

See also Balanced Repeated Replication (BRR); Correlation; Jackknife Variance Estimation; Nonresponse Bias; Simple Random Sample; Taylor Series Linearization; Variance; Variance Estimation

Further Readings

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Coverage

The term *coverage*, as used in survey research, indicates how well the sampling units included in a particular sampling frame account for a survey's defined target population. If a sampling frame does not contain all the units in the target population, then there is undercoverage of the population. If the frame contains duplicate units or other units beyond those contained in the population, then there is overcoverage. Undercoverage and overcoverage do not necessarily mean there will be coverage error associated with the frame.

Overcoverage occurs when members of the survey population are erroneously included in the survey sampling frame more than once or are included erroneously. Noncoverage (including undercoverage) occurs when members of the targeted population are erroneously excluded from the survey sampling frame. The meaning of the term *noncoverage* is not the same as the meaning of *unit nonresponse*, which is the failure to obtain complete survey data because of issues such as noncontacts, refusals, lost questionnaires, and so on.

Both overcoverage and noncoverage can occur at several junctures during the survey process. For example, in population surveys in which the sample is selected in two or more stages to obtain estimates of persons within households, coverage errors may occur at any or all stages when creating the sampling frame of primary sampling units, during field listing of housing units, or when creating a household roster of persons within a given family. Noncoverage that occurs during field listing can result if members of the survey sample are excessively expensive to locate or are part of multi-unit structures, or if maps do not accurately display the sampling area. Survey coverage is affected by the amount of time that has lapsed between obtaining the information for constructing the frame, creating the frame, drawing the sample, and finally collecting the data by methods such as personal visit, telephone, mail, Web, or by abstracting records. Several months or years may have passed during this time period, and many changes may have occurred to the units in the initial sampling frame that will not be reflected in the final sample.

Noncoverage

Noncoverage can occur when sampling units are omitted or missing from the sampling frame. For example, a sampling frame of business establishments may omit newly created businesses, or an administrative system may exclude units that failed to submit reports, or newly constructed buildings may be omitted from a housing survey. This will result in an incomplete frame from which the sample is drawn. Biases in the resulting survey estimates can occur when it is incorrectly assumed that the frame is complete or that the missing units are similar to those included in the frame, if units are actually known to be missing from the sampling frame.

A special case of noncoverage can be attributed to sampling units that are misclassified with respect to key variables of interest, such as a person's race-ethnicity or a household's vacancy status. When these key variables are missing, the sampling units cannot be properly classified in order to determine their eligibility status for the survey. In population household surveys, groups such as homeless persons or constant travelers are generally excluded from coverage. Special procedures may be necessary to account for these groups to prevent understating these populations in the survey estimates. Alternatively, if this is not feasible, it is important that published survey results document the limitations in coverage and possible errors in the survey estimates associated with imperfect coverage.

Overcoverage

Overcoverage can occur when the relationship between sampling units is not properly identified, resulting in duplicate or erroneous entries on the sampling frame. For instance, use of lists to develop the survey sampling frame might overlook events such as business mergers or changes in a facility's ownership. When the survey sampling frame is created by merging several lists, consistent identifiers for each sampling unit are essential in order to discard duplicate entries. (In practice this is very difficult to institute, and sometimes it even may require manual labor to purge all true duplicates from frames.) Potential overcoverage also occurs when sampling units cannot be identified as out of scope and are subsequently included in the survey sampling frames. Another example is in agricultural surveys, when using small grids for selecting samples of crops tends to introduce overcoverage, since many plants appear on the borderline area and field workers tend to include them; thus larger grids with smaller proportions of borderline areas are preferable for creating the survey sampling frame.

When there is overcoverage in the sampling frame due to the inclusion of out-of-scope cases, these cases may be in the sample and coded as missing during the weighting or imputation processes, if it is not possible to obtain information about them a priori so they can be excluded from the sample. This can occur in establishment surveys in which nonrespondents may be assumed to be eligible sampling units when, for instance, the establishment is no longer in operation.

Overcoverage occurs less frequently in most household surveys than noncoverage.

Solutions to Coverage Problems

It is important to routinely assess and measure survey coverage to evaluate survey quality and to improve sampling frames. For surveys in which the sample is selected in two or more stages, administering coverage rules that uniquely associate persons with households or businesses within multi-unit corporations are essential to counter both overcoverage and noncoverage. Proper training is important to verify that these rules are understood by field staff who perform tasks such as survey listing, interviewing, and providing oversight of data collection.

Typical methods to reduce or minimize coverage problems include the use of pilot tests to assess coverage; the use of multiple frames during frame construction, such as a list frame along with an area frame; the use of weighting adjustments to reduce the bias resulting from coverage errors; and truncation of the sampling frame. Pilot tests are useful for uncovering unexpected deficits in coverage and allow for survey plans to be modified in various ways.

The use of multiple frames can increase chances of selection for target population elements. To address the problem of identifying duplicate entries, one simple method is designating a principal frame for sample selection and supplementing by a frame that provides better coverage for elements that are unlikely or absent from the principal frame. This approach is taken by the U.S. Bureau of the Census, which supplements its area sampling frame (that was constructed from census information) with a list of permits for residential units built after the decennial census.

Weighting adjustments usually involve benchmarking to appropriate administrative data, so that sample estimates agree with nationally known estimates. Numerous household surveys, such as the National Survey of Family Growth in the United States, use census data in this manner.

Truncation of certain sampling units within the sampling frame is a typical compromise. The decision to truncate is made because specific sample cases, such as unregulated or smaller businesses in establishment surveys, are difficult to list. This action can help considerably to reduce both coverage problems and the cost of the survey, for example, when removal of the smaller businesses has a trivial impact on the final survey estimates. Estimates for the sampling units removed from the sampling frame may be obtained through synthetic estimation techniques, in which survey estimates are benchmarked to subgroups of the target population.

Karen E. Davis

See also Coverage Error; Frame; Noncoverage; Nonresponse; Overcoverage; Pilot Test; Sampling Frame; Target Population; Unit; Unit Coverage; Universe; Within-Unit Coverage; Within-Unit Coverage Error

Further Readings

Foreman, E. K. (1991). *Survey sampling principles.* New York: Dekker.

Gonzalez, M. (1990). *Survey coverage.* Statistical policy working paper no. 17. Retrieved December 1, 2006, from http://www.fcsm.gov/working-papers/wp17.html

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Coverage Error

Coverage error is a bias in a statistic that occurs when the target population does not coincide with the population actually sampled. The source of the coverage error may be an inadequate sampling frame or flaws in the implementation of the data collection. Coverage error results because of undercoverage and overcoverage. Undercoverage occurs when members of the target population are excluded. Overcoverage occurs when units are included erroneously. The net coverage error is the difference between the undercoverage and the overcoverage.

Bias in Descriptive and Analytical Statistics

Both undercoverage and overcoverage are biases and therefore may distort inferences based on descriptive or analytical statistics. Weaknesses in the sampling frame or the survey implementation create coverage error by compromising the random selection and thus how representative of the target population is the resulting sample. This is particularly the case if the cause of the coverage error is correlated with the characteristics being measured.

The amount of bias in descriptive statistics, such as means and totals, from undercoverage depends on the proportion of the population not covered and whether the characteristics of individuals not covered differ from those who are. If those not covered are merely a simple random sample of the population, then means will not be biased, although totals may be. For example, when estimating the mean, excluding individuals in the target population will not bias the mean if the mean of those covered equals the mean of those not covered. However, usually the exclusion of individuals is not random. More often, the excluded individuals are difficult to identify and to contact for interviews because of their characteristics. For example, a telephone survey measuring income would exclude individuals with low incomes who could not afford a telephone.

Coverage error also may affect analytical statistics, such as regression coefficients. The amount of bias in a regression coefficient from undercoverage depends on the ratio of the dependent variable's variance in the target population to that in the covered population and the quality of the fit of the regression model in the target population. If the variance of the dependent variable in the covered population is lower than the variance in the target population, the measured regression coefficient will be too small. In the telephone survey mentioned previously, the exclusion of low-income individuals would reduce the variance of income in the sampled population to be lower than in the target population. The effect on the regression coefficient is diminished when the fit of the regression model is very good in the target population.

Overcoverage also may create a bias in both descriptive and analytical statistics. The mechanism creating the bias when inappropriate or duplicate units are included mirrors the mechanism when appropriate units are excluded. The amount of bias in descriptive statistics from overcoverage depends on the proportion of the population sampled that is inappropriate and whether the characteristics of the inappropriate units differ from those in the target population. The amount of bias in a regression coefficient from overcoverage depends on the ratio of the dependent variable's variance in the target population to that in the population sampled and the quality of the fit of the regression model in the target population. Inappropriate units may cause the variance of the dependent variable to be larger or smaller than its variance in the target population.

Causes of Coverage Error

Coverage error may occur at the outset of a survey, in the sampling frame, or in the course of the survey, in the data collection. Ideally every member of the population is attached to one and only one listing record on the sampling frame. However, an exact one-to-one correspondence between population units and frame listings is often hard to find in practice. Either the frame fails to include some members of the target population, or it includes other units that are not eligible, or both.

One way to deal with a frame that is incomplete is to supplement it with a special frame or frames for the units that are not covered, resulting in what is known as a "multiple-frame" survey. For example, the researcher may have a list of all the large stores but not the small stores. Adding an area frame for sampling the small stores may be a solution to the undercoverage from the list frame.

Blanks or listings that are not members of the target population may be a problem with a sampling frame. More listings than members of the target population on a frame create overcoverage. The optimal solution for a frame with listings that are blank or not in the target population is to remove them before selecting the sample. When blanks and nonmembers can be identified during sample selection, one remedy to overcoverage is to reject such units when selected and draw another unit at random to attain the desired sample size.

Clustering of several population members into one unit on a frame may be a source of coverage error. One listing on the frame may be tied to more than one unit in the population. There are different ways that researchers still can work with the frame. One option is to take the whole cluster into the sample. The other option is to subsample within the cluster and make a weight adjustment in the estimation. For example, if the researcher wanted to interview adults but had a frame of households (e.g., in a random-digit dialing [RDD] telephone survey), the interviewer could list all the adult members of the household and then choose one member at random to interview instead of interviewing all the adults.

Multiple listings of the same individual may cause a coverage problem. When one individual in the population is attached to more than one unit on the frame, the researcher has two ways to address this problem. One is to remove the duplicate listings in the frame before selecting the sample. However, removing the duplicate listings prior to sample selection may not be practical. If the number of listings an individual has on the frame can be determined during the interview, there is another option. This option accounts for the individual's increased selection probability by weighting the unit in estimation by $1/k$ where k equals the number of times the population unit occurs on the list (such as when a person can be reached by more than one telephone number in an RDD survey).

Coverage error also may arise during the course of data collection. Interviewers need specific instructions about how to define the target population and sample unit. Otherwise, they may exclude members of the target population or include some who are not in the target population. Even experienced interviewers may have difficulties when faced with complicated situations. For example, whether a commercial structure at an address contains residential living quarters is not always clear. A business may have an apartment at the back or upstairs that is not obvious from the street. Also, an interviewer in a household survey may have to deal with ambiguities about the members of a household because a person may stay with the household only some of the time.

Longitudinal surveys that interview a sample periodically over a period of years have the potential for coverage error due to attrition, in addition to coverage concerns at the time of the initial sample selection. One approach is to estimate the attrition rate and then draw an initial sample large enough to produce a desired sample size at the end. Adjustments for the attrition may be made in the estimation.

Avoiding Coverage Error by Design

Minimizing coverage error is a major consideration when designing the survey. The measurement unit, the frame selection, and data collection and processing may contribute to coverage error if not designed properly. The researcher has to weigh many things when choosing a frame. First the list has to be available or feasible to use for sample selection. The units on the list have to be clearly defined. The extent of the coverage of the target population has to be assessed. The accuracy and completeness of the information on the list is important to assess whether the survey can be implemented without causing coverage error. Also, the amount and quality of auxiliary information on the list has to weigh on whether it will be helpful in the analysis of the data collected.

There may be more than one way to define the target population. The researcher has to assess the potential for coverage error for each way. For example, in medical expense audits, the researcher has to decide whether the units will be patients or visits to the doctor's office. In studies of income, the researcher has to decide whether the unit for measurement will be households or persons.

When selecting the units for measurement, the researcher has to be sure that those selected can answer the questions required to achieve the goals of the research. For example, using visits to doctors' offices instead of individual patients may not portray total medical expenses accurately. Also, using persons instead of households may skew the estimates of total disposable income.

Measurement of Coverage Error

Measuring coverage error is often difficult because an auxiliary data source for the target population is required. Estimates of coverage error generally cannot be made with the data collected for the survey. When a suitable auxiliary data source is available, statistics estimated with survey data may be compared to statistics estimated with the auxiliary data. Although the auxiliary data source may be available for only some of the characteristics the survey measures, such a comparison provides guidance regarding coverage error.

When using an auxiliary source for estimating coverage error, the researcher also has to be concerned about the coverage error in the auxiliary source. Even a census, which is often used to judge whether coverage error exists, may have coverage error itself. For the U.S. Population Census in 2000, two different methods estimated coverage error. Both found the net coverage error for the population overall to be very close to zero, but also found that the net coverage error rate was not uniform across the population. To illustrate the differential coverage error within groups, both methods estimated undercoverage for black males and overcoverage for nonblack females.

Compensating for Coverage Error

When auxiliary data are available for the target population, the researcher may use an adjustment to correct for coverage error. The method is a weight adjustment applied after the data are collected as opposed to corrections to the frame or methods applied during data collection to improve coverage.

A weight adjustment similar to post-stratification compensates for undercoverage, although it is sometimes used to compensate for unit nonresponse or to reduce sampling variance. After the data are collected, the sample is separated into groups for which known population totals are available and for which there may be differential coverage error. Within each group, one weighting component is applied to each member of the group. The weight for individuals in the sample equals its known group total divided by the group total estimated from the survey. The known group total may come from a census, administrative records, or other auxiliary source.

When two or more sets of marginal distributions are known, a procedure known as "raking" can be used to form the weighting adjustments in a similar way, so that estimated marginal distributions from the survey agree with each set of known marginal distributions.

Coverage Error in Surveys Using Area Frames

An area frame is constructed by dividing the geographic area of interest into mutually disjoint sections.

These sections are the units for sampling and may be areas such as counties, blocks, or districts defined for the purposes of the survey. In addition to selecting samples of housing units, area frames are often used to survey crops, wildlife, and business establishments. Area frames may be used for other topics such as a survey of school children when school districts are sample units. For example, in a multi-stage sample design, school districts could be the first-stage sample unit with the schools and students as the second- and third-stage sample units, respectively.

Area frames can have unique coverage problems when the boundaries for the sample units are ambiguous. An interviewer may have difficulty in determining whether a member of the target population is in the geographic unit selected for the sample. A tendency to include population members when the boundaries are unclear may lead to overcoverage, while the tendency to exclude members when the boundaries are uncertain may result in undercoverage.

Coverage Error in Household Surveys

The different types of household surveys have both shared concerns and unique concerns about coverage error from their frames and sampling within households for each type of survey.

In surveys of households, researchers have to be concerned not only about coverage of households but also about coverage within households (i.e., possible within-unit coverage error). Whether the survey collects data for all the household members or just some, coverage errors may occur through the interview. If the survey collects data for every member of the household, determining whom to include may be difficult because some people may have a tenuous attachment to the household.

If a survey targets only one member of the household, always interviewing the person who answers the telephone or the door may cause coverage error. Many households have one member who usually does these activities. If so, the other members of the household essentially have a zero probability of selection, which would lead to undercoverage at the person level. To achieve a random sample of respondents, the interviewers need a method for sampling within the household, which may be as simple as asking to speak to the household member with the next birthday.

Movers may be a source of coverage error, even though the frame is perfect and the sample selection and interviewing methods are perfectly designed to produce a random sample of the population. Movers may have a higher probability of selection because they may have the opportunity to be included twice, once at the old residence and once at the new residence. A survey with a long data collection period may be more vulnerable to problems with movers than one in which there is a short data collection period. Also, movers may practically have a zero probability of being selected if they are in transit while the survey is being conducted because they will be missed at both the old residence and the new residence.

People with multiple residences also may be a source of coverage error. Multiple residences are often hard to detect during interviews because some respondents tend not to report the second residence. Designing questions that allow interviewers to determine a respondent's primary residence accurately is challenging because the patterns of alternating between the residences are not uniform. Some people maintain two or more homes in different parts of the country and stay at each one several months at a time. Others commute weekly between cities, having a family home in one city and an apartment in the city where they work. These situations may cause some people to have an increased probability of selection because they would be interviewed if either of their homes were selected for the sample. Others may practically have a zero probability of selection because they would always be considered to live at the residence other than where an interviewer finds them. Interviewers need specific definitions for determining where a person lives to avoid introducing coverage errors.

Typical modes for conducting household surveys are mail, face-to-face, or telephone. Although the Internet is a fast mode of communication, no frame exists for email addresses that will provide a random sample of those who have email addresses. Of course, if such a frame existed, it would not cover those who do not have email addresses. Sometimes researchers use the Internet to gather data. In these cases, the respondents are recruited by another means that does provide a random sample and then merely convey their responses over the Internet.

Unique Coverage Error Concerns

Mail surveys use address lists as frames. A frame currently in use in the United States for mail surveys of households is the list of all the addresses where the

U.S. Postal Service delivers mail. Researchers may purchase the list from the U.S. Postal Service. No addresses are withheld if the purpose is research, although residents can request their address not be released for marketing purposes. However, such a list may have coverage problems because not every household receives mail at their houses. In addition, some people have multiple homes and thereby have a higher selection probability.

Face-to-face surveys use address lists or area frames composed of geographic areas such as blocks. When geographic areas are used for the frame, typically a list of the housing units is made in the selected areas before the interviewing begins. An interviewer starts at a particular point and proceeds around the block in the clockwise (or counterclockwise) direction, listing addresses until arriving back at the starting point. If some time has elapsed between the listing and the sample selection, new addresses may have appeared on the block. A method known as the "half-open interval" allows these new units to be linked to a unit already on the frame of addresses. When a new unit would have been listed after an address selected for the sample, the interviewer conducts an interview at the new unit in addition to the unit in the sample. The half-open interval method does not help with duplicate listings or addresses on the list for units that have been demolished or even moved, which may happen with mobile homes.

For telephone surveys of households, telephone books are not suitable for a frame because unlisted numbers, substantial in some states, are excluded. In addition, more and more people use only a cellular (mobile) telephone, and in the United States and some other countries those numbers are not included in telephone books. The method called "random-digit dialing" (RDD), which is used most often to obtain a random sample, starts with the 6-digit area code and prefix combinations that contain working residential numbers and generates telephone numbers randomly. Identifying the first 8 digits in telephone numbers with a pre-specified minimum number of telephone numbers that are listed creates the frame. In the United States, the creation of the sample starts by selecting the first 8 digits of the telephone number and then randomly generating the last 2 digits to create a 10-digit telephone number. Choosing the pre-specified minimum has to balance the trade-offs of avoiding the cost of dialing a large number of nonresidential numbers but including as many residential numbers as possible on the frame. The first 6 digits of working cellular (mobile) telephone numbers also are available in some countries.

In the United States, undercoverage from an RDD survey is possible because some telephone number banks defined by their first 8 digits will have fewer than the minimum number of listed numbers specified by the sampling design, thus giving any household in these banks a zero probability of selection. If cellular telephone numbers are excluded because of the expense, undercoverage of households that use only cellular telephones will occur. Overcoverage may also occur because many residences have more than one telephone line. To account for multiple lines, the interviewer needs to ask how many lines there are in the home. Since some lines are never answered because they are restricted to fax machines or modems, the interviewers also need to ask how many of the lines are answered. If there are k lines answered, the household's increased selection probability may be addressed by weighting the household in estimation by $1/k$, the correction for multiple listings on a frame.

One way researchers attempt to cope with the difficulty of avoiding coverage error is to recruit a group of people who agree to respond several times during a period of time, say, a year. This method usually attempts to match demographic and geographic distributions. If the recruiting is based on a random sample, then this method may be effective. However, if the recruiting is not based on random sampling, then there may be coverage error.

Coverage Error in Surveys of Events

Some surveys seek to inquire about events. There are no lists of some types of events, such as pregnancies, purchase or service of a particular product, or listening to a radio station. Some events, such as births, are recorded, but a list of such events may not be available to survey researchers for privacy reasons. The survey researcher has to rely on another type of frame to arrive at a sample of these events. Often household frames are used to sample for events. The respondents are asked if anyone in the household experienced the event during a given time period, such as within the past month. If the event is unusual, the cost of screening to find people who have experienced the event may be substantial.

Opportunities for coverage error are present because a respondent who has experienced the event may not remember exactly when it happened. The recall problem may lead to reports of events that happened prior (i.e., telescoping) to the time period or failing to report events within the time period. Undercoverage also may happen because the respondent for the screening questions may not know that the event happened to another member of the household.

Coverage Error in Establishment Surveys

Establishment surveys have their own unique sources of coverage error. Miscoding of industry, size, geographic location, or company structure may lead to frame errors that result in coverage error. The list frame may not be updated often enough to reflect the population corresponding to the survey reference period. Changes that make frames out of date include acquisitions, mergers, and growth in one line of business. In addition, the maintenance process for the list may not enter new businesses in the frame in a timely manner. Businesses that are no longer operating may remain on the list for some time after they close. There may be a delay in recording changes in a business that would cause its industry or size coding to change.

For the United States, Dun & Bradstreet has a list of businesses that is publicly available. These listings have addresses and telephone numbers. When a business has more than one location, researchers have to decide whether the target population is establishments or a more aggregated level within the company. The U.S. Census Bureau maintains its own list of businesses for its surveys, but the list is not available to the public.

Small businesses pose more difficult coverage error concerns because they are less stable than larger businesses. The process for forming the large lists is unable to keep up with the start-ups and failures in small businesses. Sometimes researchers use multiple-frame methodology that relies on a list frame and an area frame to reduce the potential for coverage error.

Mary H. Mulry

See also Area Frame; Attrition; Auxiliary Variable; Face-to-Face Interviewing; Frame; Half-Open Interval; Mail Survey; Multiple-Frame Sampling; Overcoverage; Raking; Random-Digit Dialing (RDD); Target Population; Telephone Surveys; Telescoping; Undercoverage; Unit Coverage; Within-Unit Coverage Error

Further Readings

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* New York: Wiley.

Groves, R. M. (2004). *Survey errors and survey costs.* New York: Wiley.

Kish, L. (1965/1995). *Survey sampling.* New York: Wiley.

Konschnik, C. A. (1988). Coverage error in establishment surveys. *Proceedings of the Section on Survey Research Methods* (pp. 309–314). Alexandria, VA: American Statistical Association.

Lynn, P. (1997). Sampling frame effects on the British crime survey. *Journal of the Royal Statistical Society*, *160*(2), 253–269.

Mulry, M. H. (2007). Summary of accuracy and coverage evaluation for Census 2000. *Journal of Official Statistics*, *23*(3), 345–370.

Cover Letter

A cover letter accompanies or transmits another document such as a survey questionnaire. Its purpose is to alert the respondent about the questionnaire it accompanies and to provide the details of requested actions on the part of the respondent. When used as a part of multiple communications or overall research strategy, such as an advanced contact or future reminder mailings, it can help increase response by conveying important information (e.g., research topic, survey sponsor, incentives) that is likely to influence a respondent's decision to cooperate and/or to comply fully and accurately with the survey task. As with all communications (including the questionnaire), the cover letter should be written in a way that maximizes the likelihood of participation and minimizes or eliminates any possible objectionable content.

Cover letters are an accepted and commonly used part of good survey design. There is a large amount of experimental research available on cover letter style, layout, elements, wording, and so on.

Style and Layout

Typically, a cover letter is brief (i.e., preferably one page), and it is best to print it on a formal letterhead.

Use of letterhead and stationery-quality paper speaks to the importance of the letter. Some cover letters are incorporated into a questionnaire's front cover or first page; but they usually are a separate (stand-alone) piece. When designing the cover letter text, the researcher should take into account the target population of the study and write to an educational level just below the average respondent's. For example, the language and vocabulary used in a cover letter to an organization or business or a survey of physicians or lawyers should differ from that of the general public. In writing the cover letter, one should make statements using an active voice.

The overall layout of the letter takes into consideration the chance that it will not be fully read by the respondent. One of the most important aspects is for the letter to be concise and to the point. Extensive and unneeded information will "crowd" the letter or give it a busy or daunting appearance. When composing the cover letter, one should evaluate whether information has been conveyed in other communications or on the questionnaire itself to eliminate overly redundant information, although some degree of redundancy is useful across various survey materials. The letter should incorporate the following stylistic features: (a) at least 1-inch margins on all sides, (b) indented paragraph–style, (c) either Times New Roman or Arial font, and (d) 11- or 12-point size font. There should be plenty of "white space" on the page so as to reduce respondent burden and increase the likelihood that the letter will be read.

The use of bold, underlined, or different color font can bring attention to critical pieces of information (e.g., "Once we receive your completed survey, we will send you **a $10.00 cash 'Thank You' gift**"), but should be used sparingly and for only the information most likely to increase cooperation. The style of all-capitalized font should not be used, or only minimally used, because some consider it to be "shouting" and it can be difficult to read.

Using sincere, polite wording also is highly recommended, such as the word *please* (e.g., "Please complete and return the questionnaire in the enclosed postage paid return envelope no later than May 31").

Elements

The elements listed following are used commonly in professional letters; they assume the use of common word processing and mail merge software. For specifics (i.e., number of lines between elements, left/center/right justification, etc.), see available letter or writing guides.

Date of Mailing

The date that the questionnaire is mailed is important to include. Giving no date or just month and year would be conspicuous and would fail to convey the timing of the request you are making to get the completed questionnaire returned.

Name of Addressee

Depending on the sample type and source, a name should be used to customize the letter whenever possible and appropriate. If the name of the addressee is from a third-party or matching service, it may be more beneficial not to use the name, because if the name is wrong (as it often is with matching services), the recipient may ignore the mailing even if the survey is of the residents of the mailed address, as opposed to a particular person at that address.

Address

Listing the address helps convey the personalization of the survey request. Be sure to include all relevant addressing elements to assist with accurate delivery; such as apartment number, lot, or unit number and the zip + 4 extension if available.

Salutation

The salutation greets the addressee by Dear [Mr. / Mrs. / Ms. surname]. Use of *Dear Sir* or *Dear Madam* is out of fashion. If the recipient's gender is unknown, use the full name, such as "Dear Chris Jones." If no name is available, and the survey is not one of named persons, then use a generic identifier, such as "Dear Health Survey Respondent" or even "Dear Resident."

Body of the Letter

The body of the cover letter, usually, is comprised of three to seven paragraphs and depends on the length or extent that each element is discussed. The elements of the body of the cover letter are as follows:

- *Survey Request.* The first paragraph of a cover letter serves as an introduction and conveys the key point or purpose of the mailing, that is, requesting that the respondent complete and return the enclosed questionnaire and identifying what organization is conducting the survey and why.

- *Importance of Participation.* This is a statement or even an appeal to the respondent of the importance of his or her cooperation in the research. This could include or separately state how the research results will benefit others.
- *Method of Selection.* A common concern for respondents is that they want to know how they were selected. The explanation should be worded appropriately, but succinctly, for the understanding by the target respondent (i.e., accurate but nontechnical). For example, for an RDD sample, "We used a computer to scientifically select your phone number and then compared it with publicly available records to match with this address."
- *Confidentiality.* Research has shown that including a statement of confidentiality can improve response rates. It is an ethical imperative that the researcher and sponsor organization adhere to this statement if it is pledged to a respondent.
- *Voluntary Participation.* Many research organizations or institutional review boards (IRBs) require that a statement be included to inform the respondent that their participation is voluntary.
- *Explanation of Incentive.* If an incentive is included or otherwise offered as a part of the survey, it should be mentioned in the cover letter. The researcher should consider carefully the type or amount of incentive and how it is referred to in the cover letter. A small cash incentive of a few dollars can be referred to as a "token of appreciation," consistent with social exchange theory; whereas a larger cash incentive may be referred to as a "payment for your participation" consistent with economic exchange theory.
- *Where to Get More Information.* Provide the respondent the ability to contact the researcher (i.e., mail, email, and/or toll-free telephone number).
- *Instructions for Return.* Provide any critical details about the questionnaire's return that the recipient needs or would like to know, for example, any specific instructions, return method (call-in, mail-in, and/or Internet), and the desired "return by" date.
- *Thank You.* Include a sincere sentence to thank the respondent or extend appreciation for their participation in advance of their giving it.

Complimentary Close

End the letter with a traditional close (first letter capitalized), such as, "Sincerely yours," "Yours sincerely," "Regards," "Best regards," and so on.

"Real" Signature

The complimentary close is followed by the signature, four lines down from the close, which states the writer's full name and below that her or his title. The use of an actual signature using ballpoint pen or blue ink digital signature has been found to raise response rates compared to no signature or a machine-imprinted signature. However, the use of an actual signature is judged to be impractical by most researchers when sample sizes are large. The actual (real) name of a person at the survey organization should be used, as it is unethical to use a fictitious name.

Postscript

Usually, a postscript ("P.S.") is read by the respondent. Careful consideration of what might or should be included in the postscript is important.

Charles D. Shuttles and Mildred A. Bennett

See also Advance Letter; Confidentiality; Economic Exchange Theory; Informed Consent; Leverage-Saliency Theory; Refusal Avoidance; Social Exchange Theory; Total Design Method (TDM)

Further Readings

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: Wiley.

Erdos, P. L. (1985). *Professional mail surveys* (pp. 101–117). Malabar, FL: Robert E. Krieger Publishing.

CRONBACH'S ALPHA

Cronbach's alpha is a statistic that measures the internal consistency among a set of survey items that (a) a researcher believes all measure the same construct, (b) are therefore correlated with each other, and (c) thus could be formed into some type of scale. It belongs to a wide range of reliability measures.

A reliability measure essentially tells the researcher whether a respondent would provide the same score on a variable if that variable were to be administered again (and again) to the same respondent. In survey research, the possibility of administering a certain scale twice to the same sample of respondents is quite small for many reasons: costs, timing of the research, reactivity of the cases, and so on. An alternative approach is to measure reliability in terms of internal consistency. Internal consistency would indicate that all of the items (variables) vary in the same direction and have a statistically meaningful level of correlation with each other. This can be done, for instance, using the so called split-half method. The most widespread approach, however, in the case of attitude and opinion scales, is to measure the coherence of the responses through the different items in order to discover which of the items are less correlated with the overall score: this is what item–total correlations do. A more sophisticated statistic that uses this same logic is Cronbach's alpha, which is calculated as follows:

$$\alpha = \frac{n\bar{r}}{1 + \bar{r}(n-1)},$$

where n represents the number of the items, and $\bar{r}$ is the average intercorrelation among them.

Cronbach's alpha ranges between 0 and 1. The greater the value of alpha, the more the scale is coherent and thus reliable (alpha is actually an approximation to the reliability coefficient). Some authors have proposed a critical value for alpha of 0.70, above which the researcher can be confident that the scale is reliable. The logic of this rule is that with an alpha of .70 or greater, essentially 50% (or more) of the variance is shared among the items being considered to be scaled together. Others have proposed the value of 0.75 or the stricter 0.80. If alpha is $\leq .70$, it is recommended that the scale be modified, for example, by deleting the least correlated item, until the critical value of 0.70 is finally reached or hopefully exceeded. The output of Statistical Package for the Social Sciences (SPSS) and other statistical packages used by survey researchers gives the researcher critical information on this issue, reporting the value of alpha if each of the items would be deleted. The researcher then deletes the item that, if removed, yields the highest alpha.

Since Cronbach's alpha tends to rise with the number of the items being considered for scaling, some researchers tend to solve the problem of its possible low value by building scales with numerous items. It has been noted that this praxis is often abused. In the end, a proliferation of items may yield a scale that annoys many respondents and can lead to dangerous respondent burden effects (e.g., yea-saying, false opinions, response set, satisficing).

A low value of alpha can have another explication, however. If the scale has a multi-dimensional structure (i.e., it contains more than one construct), in fact, alpha will usually be low. For this reason, alpha is not sufficient alone, because it is not a measure of unidimensionality, as some authors maintain. It would be helpful, then, before the calculation of alpha, to check for the unidimensionality of the scale through factor analysis. If two or more subsets (i.e., factors) of the scale are found, alpha should be calculated for each of the subsets separately. Therefore it is recommended that a factor analysis be conducted before calculating alpha even when alpha shows a high value, because the high value could be determined by a high correlation of the subsets, which could mask the multidimensionality of the scale. Note also that a scale can have a low value of alpha even when it is unidimensional: this can happen if there is a high random error across the data.

If alpha is negative—which is statistically possible but meaningless in interpretation—there is surely a problem in the orientation (direction) of the categories of at least some of the items being scaled. The researcher, then, has to be careful that the polarities of the items are set coherently with the concept or attitude to measure. If not, she or he needs to recode the items so that they all are scaled in the same direction.

A final matter to consider is the paradox of alpha as it approaches its maximum value (1.00). Were a scale to have an alpha of 1.00, that would mean that all items composing that scale are perfectly correlated with each other. It also would mean that any one of the items would measure the construct as well as any other of the items, and also that any one item would measure the construct as well as the entire multi-item scale. As such, if alpha values much exceed 0.90, a researcher should give consideration as to whether or not all of the items need to be measured (used) in subsequent surveys using the scale.

Alberto Trobia

See also Attitude Measurement; Opinion Questions; Reliability; Respondent Burden; Satisficing; Split-Half; Statistical Package for the Social Sciences (SPSS)

Further Readings

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York: Harper & Row.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.

CROSSLEY, ARCHIBALD (1896–1985)

Archibald Maddock Crossley was born on December 7, 1896, in Fieldsboro, New Jersey. His love for the state of his birth carried him to Princeton University in 1917; he later worked for a small advertising firm based in Philadelphia. Crossley's research career began soon afterward, in 1918, when he was asked by an executive in his firm to create a research department, something he knew nothing about. Once the department was created, Crossley began work on "Crossley Rating," which many believe is the first ratings system. Using this rating, one could estimate the number of telephone subscribers tuned in to any radio show at any given time. Creating the ratings was no easy task, requiring various Crossley aides to thumb through telephone books covering more than 80 U.S. cities. From these telephone books, researchers were able to randomly call individuals and determine to what programs they were listening. For 16 years, people were asked one by one until May 1942, when Crossley's rating system was replaced with a simpler Hooper telephone poll. Even though Crossley's measure gave no indication about what people thought of a program, it was still used to get a sense of what programs people were listening to, which soon became synonymous with good and bad programming, similar to the Nielsen and Arbitron ratings systems of today.

Crossley's work in radio ratings served as a catalyst for other research endeavors, leading him to form Crossley, Inc., in 1926, a company that still operates today under the name Crossley Surveys, created in 1954 when Crossley, Inc., merged with another firm. During this time, Crossley collaborated with George Gallup and Elmo Roper and successfully predicted the 1936 presidential election, which was made infamous in public opinion circles after the *Literary Digest* incorrectly predicted Alfred Landon would defeat Franklin D. Roosevelt, an error that Crossley and others attributed to sample bias and the misanalysis of poll returns. This experience led Crossley to participate actively in the establishment of the Market Research Council, the National Council on Public Polls, and the American Association for Public Opinion Research, for which he served as president from 1952 to 1953.

During his academic career, Crossley concentrated on the psychology of questionnaires, focusing on how question wording could affect how the intensity of a given response is measured. This led him to crusade for ethics and professional polling standards at many different levels. This in turn led him to publicly admonish the Lyndon Johnson administration in 1967 for leaking a private Crossley poll to the press in an attempt to bolster Johnson's diminishing popularity. This emphasis on the importance of research and ethics some say is Crossley's most important contribution, since it frames the way social scientists think about their research and profession. Time and time again Crossley would remind his colleagues about the importance of using public opinion research to improve the human condition. Perhaps it is appropriate that Archibald Crossley passed away in his home in Princeton on May 1, 1985, since that is where he spent the majority of his professional life. However, even in memory Archibald Crossley serves as an important reminder to all social scientists about the potential of our research and the importance of our profession.

Bryce J. Dietrich

See also American Association for Public Opinion Research (AAPOR); Ethical Principles; Gallup, George; National Council on Public Polls (NCPP); Public Opinion Research; Questionnaire Design; Roper, Elmo; Sample Design; Telephone Surveys

Further Readings

Davison, W. P. (1985). In memoriam: Archibald Maddock Crossley, 1896–1985. *Public Opinion Quarterly*, *49*, 396–397.

Moon, N. (1999). *Opinion polls: History, theory, and practice.* Manchester, UK: Manchester University Press.

Sheatsley, P. B., & Mitofsky, W. J. (Eds.). (1992). *A meeting place: The history of the American Association for Public Opinion Research.* Ann Arbor, MI: American Association for Public Opinion Research.

CROSS-SECTIONAL DATA

Cross-sectional data are data that are collected from participants at one point in time. Time is not considered one of the study variables in a cross-sectional research design. However, it is worth noting that in a cross-sectional study, all participants do not provide data at one exact moment. Even in one session, a participant will complete the questionnaire over some duration of time. Nonetheless, cross-sectional data are usually collected from respondents making up the sample within a relatively short time frame (field period). In a cross-sectional study, time is assumed to have random effect that produces only variance, not bias. In contrast, *time series data* or *longitudinal data* refers to data collected by following an individual respondent over a course of time.

The terms *cross-sectional design* and *cross-sectional survey* often are used interchangeably. Researchers typically use one-time cross-sectional survey studies to collect data that cannot be directly observed, but instead are self-reported, such as opinions, attitudes, values, and beliefs. The purpose often is to examine the characteristics of a population.

Cross-sectional data can be collected by self-administered questionnaires. Using these instruments, researchers may put a survey study together with one or more questionnaires measuring the target variable(s). A single-source cross-sectional design asks participants to provide all data about themselves with the questionnaire generally administered in a single session. A multi-source cross-sectional design gathers data from different sources, such as the sampled respondents, their supervisors, coworkers, and/or families, with different questionnaires administered to the different populations.

Cross-sectional data can also be collected by interviews. There are one-to-one interviews, panel interviews, and focus groups. In a one-to-one interview, a participant is questioned by one interviewer. In a panel interview, a participant is interviewed by a group of interviewers. In a focus group, a group of participants are simultaneously asked about their attitudes or opinions by a discussion leader or facilitator.

Cross-sectional data can be gathered from individuals, groups, organizations, countries, or other units of analysis. Because cross-sectional data are collected at one point in time, researchers typically use the data to determine the frequency distribution of certain behaviors, opinions, attitudes, or beliefs. Researchers generally use cross-sectional data to make comparisons between subgroups. Cross-sectional data can be highly efficient in testing the associations between two variables. These data are also useful in examining a research model that has been proposed on a theoretical basis. Advanced statistical tests, such as path analytic techniques, are required to test more complex associations among multiple variables. The biggest limitation of cross-section data is that they generally do not allow the testing of causal relationships, except when an experiment is embedded within a cross-sectional survey. Cross-sectional data are widely used in social science research. Some advantages in conducting cross-section studies include the following:

1. Research participants are usually more willing to cooperate in a one-time survey research study than a series of multiple surveys taken at different points in time.
2. Researchers do not need to worry about the attrition problems that often plague longitudinal studies, with some respondents not providing data at subsequent survey waves.
3. Researchers are able to collect cross-sectional data from multiple individuals, organizations, countries, or other entities.
4. Compared to longitudinal surveys, cross-sectional data are less expensive and less time consuming to gather.

However, there also are disadvantages with cross-sectional data. For example, cross-sectional data are not appropriate for examining changes over a period of time. Thus, to assess the stability of social or psychological constructs, longitudinal data are required.

Sociologists, in particular, made significant contributions to the early design and conduct of cross-sectional studies. One of the major contributors in cross-sectional design and the use of cross-sectional data was Paul Lazarsfeld. Leslie Kish made significant contributions about how to sample subjects from a target population for cross-sectional data.

Cong Liu

See also Attrition; Cross-Sectional Survey Design; Field Period; Focus Group; Interviewer; Longitudinal Studies; Sampling; Survey

Further Readings

Babbie, E. R. (1990). *Survey research methods.* Belmont, CA: Wadsworth.
Kish, L. (1965). *Survey sampling.* New York: Wiley.
Lazarsfeld, P. F. (1958). Evidence and inference in social research. *Daedalus*, *87*, 120–121.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton-Mifflin.
Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior*, *15*, 385–392.
Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In R. T. Harry & J. M. Charles (Eds.), *Handbook of research methods in social and personality psychology* (pp. 223–252). Thousand Oaks, CA: Sage.

Cross-Sectional Survey Design

A cross-sectional survey collects data to make inferences about a population of interest (universe) at one point in time. Cross-sectional surveys have been described as snapshots of the populations about which they gather data. Cross-sectional surveys may be repeated periodically; however, in a repeated cross-sectional survey, respondents to the survey at one point in time are not intentionally sampled again, although a respondent to one administration of the survey could be randomly selected for a subsequent one. Cross-sectional surveys can thus be contrasted with panel surveys, for which the individual respondents are followed over time. Panel surveys usually are conducted to measure change in the population being studied.

Types of Cross-Sectional Surveys

Cross-sectional surveys can be conducted using any mode of data collection, including telephone interviews in which landline telephones are called, telephone interviews in which cell phones are called, face-to-face interviews, mailed questionnaires, other self-administered questionnaires, electronic mail, Web data collection, or a mixture of data collection modes. A variety of sampling frames can also be used to select potential respondents for cross-sectional surveys: random-digit dialing frames, lists of addresses or (landline) telephone numbers, lists of cell phone numbers, lists of businesses or other establishments, and area probability frames. They may also use a multiple-frame approach to sampling.

Examples of cross-sectional surveys include the American Community Survey, the Decennial Census long form, and many political and opinion polls.

Design Considerations

The principles of cross-sectional survey design are those that one would normally think of for survey design in general. Designing a panel survey would be similar, except that provisions would need to be made in sampling, operations, and questionnaire design in light of the need to maintain contact with respondents and collect repeated measurements on variable of interest. Some of the considerations particular to panel surveys could apply to a cross-sectional survey that is to be repeated in the future.

The steps in designing a cross-sectional survey may be thought of as (a) conceptualization (or research design), (b) sample design, (c) questionnaire (or other data collection instrument) design, and (d) operations planning.

Conceptualization

Conceptualization includes the following:

1. Defining the study population
2. Formulating hypotheses, if any, to be tested
3. Defining the outcome (dependent) variables of interest and important classification or independent variables
4. Specifying levels of precision, such as standard errors, confidence intervals (“margins of error”), or statistical power
5. Deciding whether the survey will be repeated
6. Establishing cost limits
7. Specifying whether the nature of the data to be collected—cost or other considerations—requires a certain data collection mode

These components of the conceptualization process should define the parameters for decisions made later in the design phase, and of course can be interrelated. The researcher should also be aware that as the design progresses, some initial decisions may have to be revisited.

While the process of conceptualization occurs in designing a study, it may not always occur in a neat and orderly fashion. A researcher may be bidding in response to a request for a proposal (RFP) or have been approached by a client with a survey design in mind. In these cases, the decisions mentioned previously may have been made and not subject to much discussion, even if the researcher thinks the design could be improved considerably.

Sample Design

The sample design builds on the process of conceptualization. Steps in designing the sample include the following:

1. Selecting (or planning to construct) a sampling frame
2. Defining the strata, if any, to be employed
3. Deciding whether the sample is to be a single-stage, clustered, or multi-stage design, and
4. Determining the sample size

The sampling frame (or alternative frames) should provide adequate coverage of the study population. The nature of the frame may be determined by the study population itself, cost, or the nature of the data to be collected. In a clustered or multi-stage design, frames will be needed at each level of sample selection.

Stratification can be used to ensure proportionate representation or to allow oversampling. Multi-stage and clustered designs are usually used when the costs of data collection are high. The sample size required is a function of the parameters being estimated, the precision desired, and the expected effects on sampling error of stratification, oversampling, and clustering.

Questionnaire Design

The questionnaire design also flows from the conceptualization process. The questionnaire or other instrument translates the dependent and independent variables into specific measurements. Often, questions available from previous studies can be used or adapted; sometimes new items must be developed. Scales to measure attitudes or psychological constructs may be available from the survey research or psychological literature. New items will require cognitive testing and pretests. The form of the questions will depend in part on the mode of data collection: for example, show cards cannot be used in a telephone survey.

Other considerations in questionnaire design include the overall length of the instrument, skip patterns, and the possibility of question ordering effects.

Operations Planning

Operations planning will depend largely on the mode of data collection. Elements of the plan include staffing, scheduling, training, and monitoring.

Telephone and in-person surveys will require a staff of interviewers, supervisors, and perhaps others, such as coders, data entry personnel, and field listers. Programmers and perhaps other information systems (IS) personnel will also be needed. If the data collection is to be done by Web, or by computer-assisted telephone or in-person methods (CATI or CAPI), the IS team may play a larger role.

The schedule for the data collection can be driven by the immediacy of the needs for survey data. Relatively short data collection schedules are often called for. Cross-sectional data can be affected by seasonality and by events such as natural disasters, wars, terrorist attacks, or even something as mundane as an election or a sports event.

Training and quality control monitoring at all levels, especially of interviewers, can have a great impact on data quality.

John Hall

See also American Community Survey (ACS); Coverage; Cross-Sectional Data; Longitudinal Studies; Mode of Data Collection; Questionnaire Design; Panel Survey; Repeated Cross-Sectional Design; Sampling Frame

Further Readings

Dillman, D. (2007). *Mail and Internet surveys* (2nd ed.). Hoboken, NJ: Wiley.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* New York: Wiley.

Current Population Survey (CPS)

The Current Population Survey (CPS) is a nationally representative large-sample survey of households in the United States, conducted by the U.S. Census Bureau and cosponsored by the Bureau of Labor Statistics. The survey's chief purpose is to provide monthly labor force data, including estimates of employment and unemployment. The survey is also a rich source of data widely used by social scientists seeking descriptive population statistics about the United States. The CPS consists of a core monthly survey and special topic supplements. Each month's core survey includes demographic and employment questions. Periodic supplements cover a variety of additional topics including income, poverty, and health insurance (each March), school enrollment (each October), voting and voter registration (in November of even-numbered years), tobacco use, computer and Internet use, occupational mobility and job tenure, and other topics. Many survey methodologists and statisticians rely upon the CPS estimates as a benchmark to test the accuracy of other surveys and as a source of population statistics that form the basis for survey weights.

The CPS originated as the Sample Survey of Unemployment, administered by the Work Projects Administration in 1940. Responsibility for the survey was transferred to the Census Bureau in 1942, and revisions over the following years led the CPS to assume many of its current characteristics during the 1950s. A decades-long span of comparable measurements is available for many key operational measures. However, substantial changes were made to the CPS in 1994, including the introduction of computer-aided personal interviewing (CAPI) and computer-aided telephone interviewing (CATI) techniques.

The CPS sample consists of approximately 60,000 households each month. The survey respondent, or "reference person," provides information about each household member. Households remain in the sample for a period of 16 months and are surveyed during the first 4 months and the last 4 months of this period, with an 8-month intervening period during which they are not interviewed. One eighth of the sample is replaced with fresh sample each month, so during any given month's survey, one eighth of the sample is being interviewed for the first time, one eighth for the second time, and so on. This sample design is intended to promote continuity in month-to-month and year-to-year comparisons of estimates. In 2 consecutive months, six eighths of the sample is the same. In the same month in 2 consecutive years, half of the sample is the same. The first and last interviews are usually conducted by CAPI, and most intervening interviews are conducted by CATI.

Data collection takes place during the week containing the 19th day of the month, and questions refer to the week containing the 12th day of the month.

Response rates on the Current Population Survey have been very high. The unweighted response rate for the core monthly survey has been 90 to 93% in recent years. Response rates on the supplements are typically above 90% of those who completed the basic monthly survey, or 80 to 90% overall.

Like nearly all sample surveys of the general population, the CPS uses complex sampling procedures rather than simple random sampling. In the CPS sampling procedure, the United States is first divided geographically into approximately 2,000 primary sampling units (PSUs), which are grouped into approximately 800 strata. One PSU is chosen from within each stratum, with a probability proportional to the population of the PSU. This design dramatically reduces the cost of data collection, particularly by limiting the areas within which interviewers must travel. With this design, CPS sampling errors are somewhat larger than they would be under the impractical alternative of simple random sampling. This means that the classical approaches to hypothesis testing and the estimation of sampling error and confidence intervals (which assume simple random sampling) are not appropriate for CPS data, as these procedures would generally overstate the precision of the estimates and lead researchers to erroneously conclude that the difference between two estimates is statistically significant when it is not.

Perhaps the most widely reported estimate from the CPS is the unemployment rate. The unemployment rate measured by the CPS is the percentage of adults in the civilian labor force who are unemployed, able to work, and actively looking for work. This rate is an estimate based on a series of CPS questions about employment status and job-seeking activities. It is worth noting that the unemployment rate is not the percentage of adult Americans who are not working; that number would be lower than the unemployment rate, because the denominator in the rate is the subset of Americans who are in the labor force (i.e., those who are employed or unemployed, but excluding those who are retired or not working for other reasons). It is also notable that the sampling error in the CPS, though small, is still large enough that a month-to-month change of 0.2 percentage points or less in the unemployment rate (e.g., from 5.5% to 5.7%) is not statistically significant at the 95% confidence level. Also, like all surveys, CPS estimates are subject to nonsampling error, which should be a further reason for interpreting small differences cautiously even if they are statistically significant.

Matthew DeBell

See also Bureau of Labor Statistics (BLS); Complex Sample Surveys; Composite Estimation; Computer-Assisted

Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Rotating Panel Design; U.S. Bureau of the Census

Further Readings

U.S. Census Bureau. (2002). *Technical paper 63 revised: Current Population Survey—design and methodology.* Washington, DC: Author.

CUTOFF SAMPLING

Cutoff sampling is a sampling technique that is most often applied to highly skewed populations, such as business establishments that vary considerably in employee size, gross revenues, production volume, and so on. Data collected on establishment surveys (from businesses or other organizations, including farms) are often heavily skewed. For any variable of interest there would be a few large values, and more and more, smaller and smaller values. Therefore, most of the volume for a given data element (variable) would be covered by a small number of observations relative to the number of establishments in the universe of all such establishments. If a measure of size is used, say, number of employees or a measure of industrial capacity or some other appropriate measure, then the establishments can be ranked by that size measure. A cutoff sample would not depend upon randomization, but instead would generally select the largest establishments, those at or above a cutoff value for the chosen measure of size. This is the way cutoff sampling is generally defined, but the term has other interpretations. Four methods are discussed here.

Cutoff sampling is used in many surveys because of its cost-effectiveness. Accuracy concerns—for example, noncoverage bias from excluding part of the population—are different than in design-based sampling and are mentioned following. Note that cutoff sampling could be used for other than establishment surveys, but these are where it is generally most appropriate.

Of the following methods, the first two are probably more universally considered to be cutoff sampling:

> *Method 1.* Assign a probability of one for sample selection for any establishment with a measure of size at or above (or just above) a cutoff value, and a zero probability of selection for all establishments with a measure of size below (or at or below) that cutoff. No estimation is made for data not collected from establishments not in the sample.
>
> *Method 2.* In the second case, the same cutoff method is applied as in the first case, but estimation is made for the data not collected from establishments not in the sample.
>
> *Method 3.* A cutoff level is established, as in the first two cases, but some establishments below the cutoff are also included in the sample. This is often referred to as "take all" and "take some" stratification. An example would be a stratified random sample with a "certainty" stratum of which all members would be sampled.
>
> *Method 4.* Data may simply be collected starting with the largest establishment and through a size-ordered list of establishments until a certain point is reached by some measure or measures, possibly subjective.

Method 1 is simple and may minimize survey costs, and it may be of suitable accuracy under a couple of alternatives. First, if the main objective of a survey is to obtain information on unit prices, or some other ratio of totals, accuracy may not be a big problem. A unit price is actually the ratio of total cost to total volume of product. If each of these totals is underestimated by truncating part of the population, then the impact on the ratio of these two totals is not as adverse as it is to each of the two totals themselves. Another consideration, even for totals, may be that the data are so highly skewed that considering the smallest numbers to be zeroes may not cause an appreciable downward bias. Considering total survey error, if collecting data from more of the smallest establishments detracts from resources needed for better accuracy in collecting from the largest establishments, this may be undesirable. However, perhaps in most cases, the main impetus for Method 1 is cost-effectiveness.

Method 2 involves the use of secondary information in estimation. For example, data from administrative records may be substituted for the missing data for the excluded smaller establishments. Perhaps a better alternative would be regression model-based estimation, typically ratio estimation. This would allow for the estimation of standard errors for the totals or ratios of totals that are being estimated. To accomplish this, there must be regressor data available for every establishment, including those not in the sample. The measure of size may be one such regressor. Multiple regression may be desirable. A related method is the link relative estimator. That relates a given set of data collected between different time periods.

Method 3 is a stratified random sample design and may therefore make use of model-based, design-based, or model-assisted design-based methods, as appropriate. Estimation for Method 4 depends on the details of the application but is similar to Method 2.

For all four methods it is desirable that some thought be given to an indication of the total survey error. Cutoff sampling is often considered cost-effective, but it can also be more accurate than other alternatives if it helps to limit nonsampling error. It also generally reduces variance due to sampling error when using regression to "predict" for data not collected, but at the risk of an unknown bias. It may be argued that part of the population is not represented when a cutoff sample is applied. It is generally advisable that the likely volumes that will not be collected for key data elements should not be large compared to the inaccuracies that can be easily tolerated.

James R. Knaub, Jr.

Official Disclaimer: This is not an endorsement by the U.S. Department of Energy or the Energy Information Administration.

See also Convenience Sampling; Establishment Survey; Inference; Model-Based Estimation; Nonprobability Sampling; Nonsampling Error; Purposive Sample; Sampling Error; Stratified Sampling; Total Survey Error (TSE)

Further Readings

Bee, M., Benedetti, R., & Espa, G. (2007). *A framework for cut-off sampling in business survey design*. Discussion paper no. 9. Retrieved October 20, 2007, from http://www-econo.economia.unitn.it/new/pubblicazioni/papers/9_07_bee.pdf

Elisson, H., & Elvers, E. (2001). Cut-off sampling and estimation. *Statistics Canada International Symposium Series—Proceedings*. Retrieved March 29, 2008, from http://www.statcan.ca/english/freepub/11-522-XIE/2001001/session10/s10a.pdf

Harding, K., & Berger, A. (1971, June). A practical approach to cutoff sampling for repetitive surveys. Information Circular, IC 8516. Washington, DC: U.S. Department of the Interior, Bureau of Mines.

Knaub, J. R., Jr. (2007, April). Cutoff sampling and inference. *InterStat: Statistics on the Internet.* Retrieved May 27, 2007, from http://interstat.statjournals.net

Madow, L. H., & Madow, W. G. (1978). On link relative estimators. *Proceedings of the Survey Research Methods Section.* American Statistical Association (pp. 534–539). Retrieved February 19, 2007, from http://www.amstat.org/Sections/Srms/Proceedings

Madow, L. H., & Madow, W. G. (1979). On link relative estimators II. *Proceedings of the Survey Research Methods Section.* American Statistical Association (pp. 336–339). Retrieved February 19, 2007, from http://www.amstat.org/Sections/Srms/Proceedings

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*, 377–387.

Saerndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer-Verlag.

Sweet, E. M., & Sigman, R. S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Research Methods Section.* American Statistical Association, I (pp. 491–496). Retrieved February 19, 2007, from http://www.amstat.org/Sections/Srms/Proceedings

D

DATA MANAGEMENT

Longitudinal projects and other large surveys generate large, complex data files on thousands of persons that researchers must effectively manage. The preferred data management strategy for such large, complex survey research projects is an integrated database facility built around modern relational databases. If one is dealing with a relatively small, simple questionnaire, many carefully implemented methods for data collection and data management will work. What needs to be done for technologically complex surveys touches upon all the considerations that can be given to less complex survey data sets.

As the scale, scope, and complexity of a survey project grow, researchers need to plan carefully for the questionnaire, how the survey collects the data, the management of the data it produces, and making the resultant data readily available for analysis. For these steps to run smoothly and flow smoothly from one to the other, they need to be integrated. For these reasons, relational database management systems (RDBMS) are effective tools for achieving this integration. It is essential that the data file preserve the relationships among the various questions and among the questionnaire, respondent answers, the sampling structure, and respondent relationships. In birth cohort or household panel studies there are often complex relationships among persons from the same family structure or household. In longitudinal surveys there are also complex relationships among the answers in various waves that result from *pre-fills* (i.e., data carried forward) from previous surveys and bounded interviewing techniques that create event histories by integrating lines of inquiry over multiple rounds of interviewing. Failure to use an RDBMS strategy for a longitudinal survey can be considered a serious error that increases administrative costs, but not using RDBMS methods in large and complex cross-sectional surveys can be considered just as big an error.

Structure

Questionnaires often collect lists, or rosters, of people, employers, insurance plans, medical providers, and so on and then cycle through these lists asking sets of questions about each person, employer, insurance plan, or medical provider in the roster. These sets of related answers to survey questions constitute some of the tables of a larger relational database in which the connections among the tables are defined by the design of the questionnaire. One can think of each question in a survey as a row within a table, with a variety of attributes that are linked in a flexible manner with other tables. The attributes (or columns) within a question table would contain, at a minimum, the following:

- The question identifier and the title(s) associated with the variable representing the question's answer with the facility to connect the same question asked in

different sweeps or rounds of a longitudinal survey. This same facility is useful in repeated cross-sections.
- Descriptors that characterize or index the content of the question (alcohol use, income, etc.).
- The question text.
- A set of questions or check items that leads into the question (in practice this information is contained in the skip patterns of contingency questions).
- A set of allowable responses to the question and data specifications for these allowable responses (whether the answer is a date, time, integer, dollar value, textual response, or a numerical value assigned to a categorical response, such as 1 = Yes, 0 = No).
- For multi-lingual surveys, there would be separate tables for question text and pick-lists for each language. This greatly simplifies the preparation and management of different survey versions for different languages that share the same core structure.
- Routing instructions to the next question, including branching conditions driven by the response to the current question, or complex check items that are contingent on the response to the current question as well as previous responses.
- Real-time edit specifications imposed upon dates, currency amounts, and other numerical (i.e., non-pick-list) data, such as numerical values that require interviewer confirmation (soft range checks) or limits on permissible values (hard range checks).
- Pre-loaded values.
- Text fill specifications.
- Instructions to assist the interviewer and respondent in completing the question and/or show cards, audio files used for audio computer-assisted self-interviews.
- Date and time stamps for the question, indicators of multiple passes through the question, and time spent in the question (this preserves an audit trail for each step in the questionnaire).
- Archival comments about the accuracy or interpretation of the item or its source or "See also notes" referring the user to associated variables that are available to users in the data set.
- Notes to the support staff about complexities associated with the question to document the internal operation of the survey.
- Links to supporting documentation produced by the survey organization or, in the case of standard scales or psychometric items, a URL to more comprehensive documentation on the item.

These attributes of questions often are referred to as "metadata." With RDBMS methods these pieces of information that describe a question are automatically connected to the variables generated by that question. For example, metadata include which questions lead into a particular question and questions to which that question branches. These linkages define the flow of control or skip pattern in a questionnaire. With a sophisticated set of table definitions that describes virtually any questionnaire, one can "join" tables and rapidly create reports that are codebooks, questionnaires, and other traditional pieces of survey documentation. The questionnaire itself is not "programmed" but rather is formed by the successive display on the screen of the question's characteristics, with the next question determined either by direct branching or by the execution of internal check items that are themselves specified in the question records. Sequential queries to the instrument database display the questions using an executable that does not change across surveys but guides the interview process through successive question records.

By breaking down the survey into a sequence of discrete transactions (questions, check items, looping instructions, data storage commands, etc.) stored in a relational database, with each transaction being a row in a database table and the table having a set of attributes as defined in the relational database, one can efficiently manage survey content, survey data, data documentation, and even public user data extraction from a single integrated database structure.

Web Integration

When the tools that reference the master database are Web enabled, staff at any field organization in the world can access this resource and share it. Access control and security measures are necessary, of course. Some users can be given access to some parts of the data set with varying read/write permissions. One person might only be able to edit database fields related to documentation and so on.

When the data capture system is built for the Web, multi-modal surveys on the Web (including cell phone Internet connections), computer-assisted telephone interview (CATI), or computer-assisted personal interviewing (CAPI) become simple to execute. (CAPI is done either by putting a client and server on the laptop or tapping into the cellular network with a wireless modem and using the Web.) The organizations involved in survey data collection are increasingly keen on multi-modal surveys in order to accommodate difficult users who have very particular preferences about how they want to do the interview. This technology meets that need.

Software

Relational database software is a major software industry segment, with vendors such as Oracle, Sybase, IBM, and Microsoft offering competitive products. Many commercial applications use relational database systems (inventory control; accounting systems; Web-based retailing; administrative records systems in hospitals, welfare agencies, and so forth, to mention a few), so social scientists can piggyback on a mature software market. Seen in the context of relational databases, some of the suggested standards for codebooks and for documenting survey data, such as the data documentation initiative (DDI), are similar to relational database designs but fail to use these existing professional tool sets and their standard programming conventions. Superimposing a DDI structure for documentation also fails to make an organic connection among the management of the instrument, management of the data, and the dissemination of the data. Rather than including the questionnaire specification in an RDBMS at the outset, the DDI approach requires the instrument to be retrofitted into DDI form with additional labor time and its attendant costs and fails to exploit the economies of scope RDBMS methods provide.

Either one plans for survey complexity at the outset of the effort or one retrofits the data from the field into an RDBMS, which amounts to paying for the same work twice or three times because of all the steps taken to manage these projects. For example, the designers must write down the questionnaire specifications. This sounds simple, but it is virtually always the case that the document the design team produces does not cover every contingency that can occur and where the instrument must branch in that case. For example, one needs to specify not only what is to happen if the respondent refuses to answer each question or says, "I don't know"; one must also decide how to handle any internal check item that encounters an answer with an item nonresponse. This means the questionnaire programmer needs to go back and forth with the design team to ensure the instrument is faithful to their intentions. Once designed, the instrument must be tested, and one needs a testing protocol that can test out the many pathways through the instrument, especially the unintended pathways. After the data are collected, they come back to the central office, but in what form? How are these data documented? How are the data checked during the field period to intercept serious problems before they affect too many cases? And then how are the data relayed to the documentation system? Every time the data or instrument changes hands, misunderstandings and errors are likely to occur. The best protection against this sort of human error is to keep a single integrated archival system that every step of the process references and uses.

The primary data collector has several data management choices:

1. Design the entire data collection strategy around a relational database that integrates with the design and testing process and also integrates with the data dissemination and documentation process that generates exports to SAS, Statistical Package for the Social Sciences (SPSS), STATA, and so on.
2. Take questionnaire specifications and program the instrument into some system, iteratively test and correct, migrate the post-field data and instrument information into a relational database for archiving, and then release the data in ASCII with documentation materials developed and maintained separately. One would produce control commands that allow SAS, SPSS, STATA, or a similar package to read the ASCII data. Alternatively, the data could be released as SAS, SPSS, or STATA system files accepting the very limited documentation tools they provide.
3. Follow #2, but without a relational database as the archival tool and try to manage the linkages with some other system, possibly a statistical software package that strips out most of the metadata implicitly present in the data capture software.

SAS, SPSS, and STATA are effective statistical packages, and one can move data among them with a package like STATA's Stat/Transfer. Statistical packages are themselves starting to incorporate relational database features. For example, SAS supports standard query language (SQL) queries to relational databases, and it also connects to relational databases. This means that building the project architecture around an RDBMS is entirely consistent with the use of established statistical packages for a wide variety of analytic and survey support activities. The trend for many years has been toward relational databases to manage databases. These tools were originally focused on large enterprise-level data management problems, but their strengths have led to their diffusion to a wider array of applications. When setting up large survey research projects, social scientists may benefit from building their data management strategies and staff resources around relational database management systems.

Randall J. Olsen

See also Codebook; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Contingency Question; Event History Calendar; Longitudinal Studies; Metadata; Multi-Mode Surveys; Panel Survey; Repeated Cross-Sectional Design; SAS; Statistical Package for the Social Sciences (SPSS); STATA; Wave

Further Readings

Elmasri, R. A., & Navathe, S. B. (2001). *Fundamentals of database systems.* New York: Addison-Wesley.

Gray, J., & Reuter, A. (1992). *Transaction processing: Concepts and techniques.* San Francisco: Morgan Kaufmann.

Kroenke, D. M. (2001). *Database processing: Fundamentals, design and implementation.* Upper Saddle River, NJ: Prentice Hall.

Stern, J., Stackowiack, R., & Greenwald, R. (2001). *Oracle essentials: Oracle9i, Oracle8i and Oracle 8.* Sebastopol, CA: O'Reilly.

Data Swapping

Data swapping, first introduced by Tore Dalenius and Steven Reiss in the late 1970s, is a perturbation method used for statistical disclosure control. The objective of data swapping is to reduce the risk that anyone can identify a respondent and his or her responses to questionnaire items by examining publicly released microdata or tables while preserving the amount of data and its usefulness.

In general, the data swapping approach is implemented by creating pairs of records with similar attributes and then interchanging identifying or sensitive data values among the pairs. For a simplistic example, suppose two survey respondents form a "swapping pair" by having the same age. Suppose income categories are highly identifiable and are swapped to reduce the chance of data disclosure. The first respondent makes between \$50,000 and \$60,000 annually, and the other makes between \$40,000 and \$50,000. After swapping, the first respondent is assigned the income category of \$40,000 to \$50,000, and the second respondent is assigned \$50,000 to \$60,000.

One benefit of data swapping is that it maintains the unweighted univariate distribution of each variable that is swapped. However, bias is introduced in univariate distributions if the sampling weights are different between the records of each swapping pair. One can imagine the impact on summaries of income categories if, in the example given, one survey respondent has a weight of 1, while the other has a weight of 1,000. A well-designed swapping approach incorporates the sampling weights into the swapping algorithm in order to limit the swapping impact on univariate and multivariate statistics.

There are several variations of data swapping, including (a) directed swapping, (b) random swapping, and (c) rank swapping.

Directed swapping is a nonrandom approach in which records are handpicked for swapping. For instance, a record can be identified as having a high risk of disclosure, perhaps as determined through a matching operation with an external file, and then chosen for swapping. *Random swapping* occurs when all data records are given a probability of selection and then a sample is selected using a random approach. The sampling can be done using any approach, including simple random sampling, probability proportionate to size sampling, stratified random sampling, and so on. Once the target records are selected, a swapping partner is found with similar attributes. The goal is to add uncertainty to all data records, not just those that can be identified as having a high risk of disclosure, since there is a chance that not all high-risk records identified for directed swapping cover all possible high-risk situations. Finally, *rank swapping* is a similar method that involves the creation of pairs that do not exactly match on the selected characteristics but are close in the ranking of the characteristics. This approach was developed for swapping continuous variables.

The complexities of sample surveys add to the challenge of maintaining the balance of reducing disclosure risk and maintaining data quality. Multi-stage sample designs with questionnaires at more than one level (i.e., prisons, inmates) give rise to hierarchical data releases that may require identity protection for each file. Longitudinal studies sometimes involve adding new samples and/or new data items over the course of several data collections. Data swapping may be incorporated in longitudinal studies to ensure that all newly collected data are protected. Also in survey sampling, data-swapping strategies incorporate sampling weights by forming swapping partners that minimize or reduce the amount of bias introduced through the swapping process.

Another aspect of data swapping to be emphasized is that careful attention is needed for maintaining data consistency. Surveys typically contain highly related variables, skip patterns, or multiple response items (i.e., "Check all that apply"). When any one data item is swapped, all items directly related to the swapped item must be swapped as well; otherwise data inconsistencies will be created.

The amount of swapping conducted, as determined by the swapping rate, is designed to protect the confidentiality of the data without affecting its usability. There is no established literature on determining swapping rates. In practice, the threat of a "data snooper" using other publicly available data impacts the swapping rate as well as whether some of the data are unique. When data swapping is conducted, the swapping approach can be tested and the impact evaluated. If it is determined that the integrity of the data is violated, then the swapping parameters can be modified and reprocessed. Last, in order to ensure that the confidentiality edits are not reversible, the swapping rate and the swapping variables are typically not revealed.

Thomas Krenzke

See also Confidentiality; Disclosure Limitation; Perturbation Methods

Further Readings

Dalenius, T., & Reiss, S. P. (1978). Data-swapping: A Technique for disclosure control (extended abstract). In *Proceedings of the Section on Survey Research Methods* (pp. 191–194). Washington, DC: American Statistical Association.

Fienberg, S., & McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases*. Lecture Notes in Computer Science (Vol. 3050, pp. 14–29). Berlin/Heidelberg: Springer.

Moore, R. A. (1996). *Controlled data-swapping techniques for masking public use microdata sets.* Statistical Research Division Report Series, RR96-04. Washington, DC: U.S. Bureau of the Census.

Reiss, S. P. (1984). Practical data-swapping: The first steps. *ACM Transactions on Database Systems*, *9*, 20–37.

Debriefing

Debriefing in survey research has two separate meanings. It is used to refer to the process whereby qualitative feedback is sought from the interviewers and/or respondents about interviews conducted and surrounding survey processes. It also is used to refer to the process whereby "justified" deception has been used by the researchers, and, following ethical research practices, respondents are then debriefed after the study ends to explain the deception to them and try to undo any harm that may have been caused by the deception.

Debriefing to Gain Qualitative Feedback

Debriefings for the purpose of gaining qualitative feedback occur in three critical phases:

1. During survey development
2. Ongoing during survey administration
3. Upon survey completion

Debriefings during survey development are the most common and the most valuable. In such debriefings, information is sought on issues that prove difficult for either interviewer or respondent, with the aim of improving the survey instruments, survey protocols, and/or interviewer training materials. The relative emphasis will depend on what other survey development activities have been undertaken; for example, respondent interpretation of questions and requests for clarification will be given less weight in a debriefing if a full cognitive interviewing process preceded the pilot test.

It is less common for a debriefing to occur during the main phase of interviewing; however, such debriefings are valuable to allow for fine-tuning of processes, answer categories, or interpretation of data. Generally it is not desirable to change any questions, as that will preclude the standardization usually sought; however, it may be appropriate to add clarifying transitional phrases in the questionnaire or clarifying questions at the end of the questionnaire if mid-survey debriefings identify serious issues that were not detected during the development phase.

Debriefings following a survey usually focus on the interpretation and limitations of the data collected. Debriefings involving respondents may also include an element of benchmarking or comparison, with information fed back to the respondent on how his or her responses compared with others surveyed. This may be for either the survey sponsor's benefit (particularly with business surveys, increased cooperation can often be obtained by the promise of such data, as long as confidentiality pledges are honored), or for the respondent's benefit (as may be the case if the survey is part of an audit procedure).

Techniques Used for Qualitative Informational Debriefings

Focus group techniques are the most often employed for interviewer debriefings, with the interviewers gathered together so that observations by one can be validated (or not) by the group. As with all focus groups, a skilled moderator is needed to balance the contributions of the participants, to keep the discussion on track, and to correctly interpret the information gathered in the discussion, so that forceful opinions are not misinterpreted as fact, and conclusions are considered within the context of the motivations of the participants. Often interviewers will be asked to complete a debriefing questionnaire prior to the focus group, to help them prepare for the discussion and/or to provide additional data for later analysis.

One-on-one interviews are more commonly used for respondent debriefings, particularly where the debriefing is a variation on cognitive interviewing techniques aimed at uncovering the various interpretations of the questions and the perceived meanings of various answer categories.

As useful as debriefing material is, at the development stage it should always complement, not replace, analysis of data collected during the pilot test. Such analysis should include at a minimum:

- Operational costs (call records, travel records, pay claims)
- Distribution of responses to questions over answer categories, compared across respondent groups and across interviewers
- Examination of responses given to open-ended questions

Such analysis can identify areas to focus on during the debriefing process and afterward to test hypotheses formed during the debriefing.

Debriefings Associated With Deception in Research

There are times when survey researchers are justified in using deception as part of their research design; for example, the need to keep respondents blind to the "real" purpose of a study until after all data have been gathered for the study. Doing so could be justified if the respondents' answers would be influenced (biased) if they understood the real purpose before their data were gathered. In these instances, it is the ethical responsibility of the researchers to debrief all respondents about the deception. This could be done in person, via telephone, via mail, and/or via an email, depending on the appropriateness of the mode of debriefing in light of the nature and extent of the deception. Through the debriefing process the researchers would (a) inform the respondents of the deception, (b) explain why it was used, (c) provide some opportunities for respondents to express any concerns they had with the deception, and (d) try to undo any harm the deception may have caused any respondent. (Sometimes, undoing the harm that deception in research causes is a very complicated, long-term, and expensive proposition.) In some instances with deception, researchers may need to gather quantitative data on the possible harm the deception may have caused as part of the debriefing of respondents, above and beyond any informal qualitative opportunities provided to respondents to express their concerns about the deception in the debriefing.

Jenny Kelly and Paul J. Lavrakas

See also Cognitive Interviewing; Deception; Ethical Principles; Focus Group; Pilot Test

DECEPTION

According to *Webster's Dictionary*, *deception* is the act of making a person believe what is not true; that is, misleading someone. The use of deception in survey research varies in degree. Typically, its use by researchers is mild and is thought to cause no harm to survey respondents and other research subjects. At times, however, the use of deception has been extremely harmful to research subjects. Thus the nature of deception involved in research must be carefully considered. Currently, contemporary researchers in the academic and government sectors submit research proposals to their institutional review board (IRB) primarily to ensure that research participants are protected from harm. In the commercial sector in the United States, this process may not be followed as closely.

It is not uncommon in survey research that some deception occurs, especially in the form of not telling respondents in advance of data collection what is the actual purpose of the study being conducted. The justification for this type of deception is the fact that telling respondents of the actual study purpose in advance of gathering data from them is likely to bias their responses.

For example, psychologists studying differences in thought patterns of depressed and nondepressed individuals may use mild deception in the form of omission of information to avoid sensitizing the subjects to the purpose of the study and thereby biasing the findings. For example, one study conducted by Carla Scanlan in 2000 did not disclose to subjects that the purpose of administering a particular screening questionnaire was to identify depressed and nondepressed subjects; the questionnaire was an untitled version of the Beck Depression Inventory—II (BDI-II), which asked subjects to read 21 sets of statements and choose the statement in each set that best described how she or he had been feeling for the past 2 weeks, including today. The consent form merely stated that the participant would fill out various questionnaires in order to determine for which experiments subjects qualified. Later, subjects were told that the purpose of this particular research project was to study the emotional state of students coming to college for the first time. After data collection and data analysis were completed, a written summary of the results was provided to those interested in the outcome. This debriefing process was complete and disclosed the purposes of the research. If the purpose of the research had been fully disclosed to participants beforehand, data collection would have been compromised.

In another example, in 2006, Scott Keeter conducted several studies in order to investigate whether cell phone only individuals differed from individuals who had landlines. That goal was not disclosed at the outset of the call; some of the questions were political in nature and others were demographic. The purpose of the call was given as a political survey, although the real intent was to investigate how cell only individuals differed from landline users. In this example, failing to disclose this purpose harmed no one and preserved the integrity of the survey responses, and it was deemed that no debriefing was necessary.

Although the uses of mild deception in survey research almost never causes harm to the respondent, there have been nonsurvey research situations utilizing deception that have caused grievous harm to the participant. For instance, the infamous Tuskegee Syphilis Study was conducted from 1932 to 1972 in Macon County, Alabama. The purpose of this study was to investigate the progression of untreated syphilis. The men (all blacks) were told that they were receiving treatment for their disease when actually it was actively withheld; the researchers secured the cooperation of all medical personnel in the county to withhold treatment from the men. Although penicillin became the standard treatment for syphilis in 1947, it continued to be withheld from the participants in the Tuskegee Syphilis Study until 1972. Some of the men had untreated syphilis infections for 40 years before they finally received treatment, but, shamefully, many of the men did not survive the disease. By 1947, if not earlier, their suffering and deaths could have been easily prevented by a penicillin injection. No one ever told them. In this case, research deception caused irreparable harm and death.

During recent presidential election years, a form of "survey" has been carried out that pretends to be gathering opinions from potential voters but in fact is an attempt to sway large numbers of voters' opinions in a particular direction as a primary approaches. This practice is known to survey professionals as a *push poll* and is actually a form of political telemarketing. For example, members of an organization that support Candidate X hire personnel to stage a telephone "survey" in which initially it may appear that a legitimate survey is being conducted. However, after the apparent legitimate start of the "interview," the person administering the "survey" begins to convey unfavorable and often false information about Candidate Y in the guise of survey questions. This is done to persuade the person being "interviewed" to vote against

Candidate Y. No debriefing takes place in these push polls, and the deceptive practice is highly unethical.

In contrast, if this approach were being done as part of a legitimate survey that involved deception, at the conclusion of the interview an ethical researcher would have interviewers debrief the respondents about the deception that took place. For example, the debriefing would honestly disclose why the false information was conveyed about Candidate Y and a sincere attempt would be made to undo any harm that the deception may have caused, including informing the respondent that the information about Candidate Y in the questions was not accurate.

Carla R. Scanlan

See also Debriefing; Disclosure; Ethical Principles; Institutional Review Board (IRB); Protection of Human Subjects; Pseudo-Polls; Push Polls

Further Readings

American Association for Public Opinion Research. (2007, June). *AAPOR statement on "push" polls.* Retrieved March 29, 2008, from http://www.aapor.org/aaporstatementonpushpolls

American Psychological Association, Ethics Office: http://www.apa.org/ethics

Keeter, S. (2007, June 20). *How serious is polling's cell-only problem?* Washington, DC: Pew Research Center. Retrieved March 29, 2008, from http://pewresearch.org/pubs/515/polling-cell-only-problem

Mayo Clinic. (2006, October 27). *Syphilis*. Retrieved March 29, 2008, from http://www.mayoclinic.com/health/syphilis/DS00374

Scanlan, C. R. (2000). *An investigation of the effect of writing about traumatic events on knowledge structures in dysphoric individuals.* Unpublished doctoral dissertation, Ohio University, Athens.

Deliberative Poll

A deliberative poll is a methodology for measuring public preferences that combines small group discussions and traditional scientific polling. It was created by James Fishkin, political science and communications professor, with the goal of improving the quality of public opinion expression and measurement.

Fishkin argues that traditional polls often do not provide good measures of public opinion because members of the public are not knowledgeable enough about the important issues of the day and do not have the motivation or opportunity to engage in deliberation on the issues. He first proposed the idea of deliberative polling in 1988 as a corrective. Fishkin, who has since trademarked the term *Deliberative Poll*, currently conducts deliberative polls through the Center for Deliberative Democracy at Stanford University.

Typical deliberative polls have three main stages. First, a traditional public opinion poll is conducted of the population of interest, for example, all voting-age adults in the United States. A probability sample of this population is selected and respondents, who agree to participate in all the stages, are asked standard survey questions on selected issues along with some background and demographic questions. Respondents are then sent briefing materials that provide information about these same issues. In the second stage, respondents travel to a given location to deliberate on these issues. The deliberations take the form of small group discussions and can include sessions where participants are able to question experts. Some more recent deliberative polls have used online deliberations. In the third stage, the participants are interviewed again using traditional survey techniques to see whether their views changed as a result of their deliberative participation. Fishkin's view is that this second survey shows what public opinion would look like if the entire population were more informed and able to engage in deliberations on these issues.

The first national deliberative poll in the United States (called the National Issues Convention) was conducted in Austin, Texas, in January 1996, at a cost of about $4 million. A second National Issues Convention was conducted in Philadelphia, Pennsylvania, in January 2003, which was followed by the first online deliberative poll. Some utility companies in the United States have also used deliberative polling at the local level to get public input on energy policies. Deliberative polls have also been conducted internationally in such countries as Australia, Britain, Bulgaria, China, Denmark, Greece, Italy, and Northern Ireland.

Some public opinion researchers have raised scientific concerns about deliberative polling. One challenge is getting a representative sample of survey respondents to participate in the deliberations. In the 1996 National Issues Convention, older respondents, those with less education, and the less politically active were less likely to travel to Austin for the weekend of deliberations. However, selection differences were less prevalent on the issue questions.

Another concern is whether group discussions are the best approach for disseminating information. Deliberative poll participants generally take the group discussion task seriously, but criticisms have been raised about the quality of the discussions and the accuracy of information exchanged in them. A related criticism of the discussions is the potential impact of group dynamics. In group situations, people can be influenced by normative factors unrelated to the strength or merits of the arguments. In addition, differences in discussion participation rates can also have an impact on opinions. Not everyone is equally motivated or has the same ability to participate in group discussions. The more vocal and persuasive members of the group may have a disproportionate influence on the outcome of the deliberative poll.

There also has been debate about the amount of opinion change that is produced by deliberative polling. For example, in the 1996 National Issue Convention, Fishkin pointed to a number of statistically significant shifts in aggregate opinion as a result of participation in that deliberative poll. Other researchers have argued that there were relatively few meaningful changes in aggregate opinion after this significant effort to educate members of the public and have them participate in extensive deliberations. This was taken as evidence of the robustness of public opinion as measured by traditional public opinion polls that can be conducted at a fraction of the cost of a project like the National Issues Convention Deliberative Poll. Larger shifts in aggregate opinion have been found, for example, in deliberative polls conducted for utility companies on esoteric issues for which opinions are weakly held or nonexistent and public interest and knowledge are very low.

Daniel M. Merkle

See also Focus Group; Poll; Public Opinion

Further Readings

Center for Deliberative Democracy, Stanford University: http://cdd.stanford.edu

Fishkin, J. (1995). *The voice of the people: Public opinion and democracy.* New Haven, CT: Yale University Press.

Merkle, D. M. (1996). The National Issues Convention deliberative poll. *Public Opinion Quarterly*, *60*, 588–619.

Demographic Measure

Demographic measures are questions that allow pollsters and other survey researchers to identify nonopinion characteristics of a respondent, such as age, race, and educational attainment. Demographic measures typically are used to identify key respondent characteristics that might influence opinion and/or are correlated with behaviors and experiences. These questions are usually found at the end of a questionnaire. Reasons for this are (a) to engage or otherwise build rapport with the respondent by asking substantive questions of interest earlier in the questionnaire; (b) to lessen the likelihood that asking these personal questions will lead to a refusal to continue completing the questionnaire (i.e., a breakoff); (c) to prevent priming the respondent; and (d) to allow the respondent to answer the core questions before possibly boring him or her with the mundane demographic details.

Demographic measures are important because numerous studies have demonstrated that opinions are formed primarily through an individual's environment. This environment socializes us to think and behave in accordance with community norms and standards. As a result, by identifying these demographic measures, pollsters are better suited to understand the nature of public opinion and possibly how it might be formed and modified.

Demographic measures are also very important because they allow researchers to know how closely the sample resembles the target population. In a national sample of U.S. citizens, for example, researchers know what the population looks like, demographically, because the federal government conducts a census every 10 years and updates those data annually thereafter until the next census. As such, researchers know the percentages of the population based on race, gender, age, education, and a whole host of other demographic characteristics. A simple random sample of the population ideally should resemble the population, and demographic measures allow researchers to see how well it does. For example, because survey nonresponse often correlates with educational attainment, most surveys of the public gather data from proportionally far too many respondents who earned college degrees and far too few respondents who did not graduate from high school. Knowing the demographic characteristics of the sample respondents (in this case, educational attainment) allows the researchers to adjust (weight) their sample to the known population characteristics. This can be done with greater confidence and accuracy if the wording of the demographic question in the survey matches the wording of the question for the same characteristics that was used to produce the universe estimates (e.g., the wording used by the U.S. Census).

The length of the questionnaire often limits the number of demographic questions asked. Accordingly, demographic measures must be carefully selected to best allow further analysis. There are a number of standard demographic questions that are nearly always asked, including questions about age, gender, income, race, Hispanic ethnicity, and education. Questions designed to identify these characteristics have become fairly standardized and often follow the ways the federal government gathers these data in the census and/or other surveys they conduct. Other common demographic measures identify the respondent's political party, political ideology, marital status, religious preference, church attendance, voter registration status, geographic place of residence, and number of children. Occasionally, the nature of a poll or other survey might cause specific other demographic questions to be asked, such as questions about military service, union membership, sexual orientation, type of employment, type of housing unit, and years lived in one's neighborhood.

These demographic measures also allow for simple breakdowns of the survey results into subgroups. Although it might be nice to know that 48% of the country approves of the job the president is doing, it may well be more informative to know that 88% of Republicans and 15% of Democrats approve of the president's job performance. Regardless of the purpose of the questionnaire, demographic measures provide a clearer picture of public preferences, dispositions, behaviors, and experiences. For instance, a marketing firm might find that men between the ages of 30 and 40 are the most likely to use a particular product. Marketers can then use this information to design advertisements that would appeal to that particular group. In short, demographic measures allow for a more nuanced understanding of the public by allowing researchers to examine the details that are absent at the aggregate level by filling in the background information.

James W. Stoutenborough

See also Census; Opinions; Poll; Population; Pollster; Questionnaire; Random Sampling; Respondent; Simple Random Sample; Weighting

Further Readings

Haines, M. R., & Steckel, R. H. (Ed.). (2000). *Population history of North America.* Cambridge, UK: Cambridge University Press.

Murdock, S. H., Kelley, C., Jordan, J., Pecotte, B., & Luedke, A. (2006). *Demographics: A guide to methods and data sources for media, business, and government.* Boulder, CO: Paradigm Press.

Dependent Interviewing

Dependent interviewing is a method of scripting computer-assisted survey questionnaires, in which information about each respondent known prior to the interview is used to determine question routing and wording. This method of personalizing questionnaires can be used to reduce respondent burden and measurement error. The prior information can be incorporated reactively, for in-interview edit checks, or proactively, to remind respondents of previous answers.

Dependent interviewing exploits the potential of scripting computer-assisted questionnaires such that each interview is automatically tailored to the respondent's situation. This can be done using routing instructions and text fills, such that both the selection of questions and their wording are adapted to the respondent's situation. Both routing and text fills are usually based on responses to earlier questions in the questionnaire. Dependent interviewing in addition draws on information known to the survey organization about the respondent prior to the interview. In panel surveys, where dependent interviewing is mainly used, this information stems from previous waves of data collections. For each panel wave, prior survey responses are exported and stored together with identifying information (such as name, address, and date of birth) used by interviewers to locate sample members eligible for the round of interviewing.

The previous information can be incorporated into the questionnaire script to reduce respondent burden and measurement error. In panel surveys, a set of core questions are repeated at every interview. For respondents whose situation has not changed between interviews, it can be frustrating and lengthen the interview unnecessarily to have to answer the same questions repeatedly. With dependent interviewing, information from previous waves can be used to verify whether a respondent's situation has changed. If not, and if the responses given in the previous interview still accurately reflect the respondent's situation, the questionnaire script can automatically route the respondent around unnecessary redundant questions. Responses from previous waves can then be filled in for the current wave. For open-ended questions such as those regarding occupation, this not only reduces the length of the interview, but also of coding time.

In general, the purpose of asking the same questions at different points in time is to generate data that can be used to investigate individual-level change. Estimates of change from panel surveys, however, tend to be biased. This is because responses about the reference period reported in one interview tend to be internally consistent but are not necessarily consistent with responses given in earlier interviews. These longitudinal inconsistencies can be due to respondent errors (such as simple variation in the way the respondent understands a question or describes her or his situation, recall errors, or estimation strategies used to compute responses), or interviewer errors, coding errors, or processing errors.

A consequence of these inconsistencies is the phenomenon called the "seam effect." Dependent interviewing can be used to remind respondents of previous responses or for edit checks to verify whether apparent changes are true. The hope is that this will reduce response variance, improve respondent recall, and catch interviewer errors. Routing around redundant open-ended questions and imputing codes from previous waves further increases longitudinal consistency. Dependent interviewing has been shown to effectively reduce, although not completely eliminate, seam effects.

The prior information can be incorporated into the questionnaire in one of two ways: (1) reactively or (2) proactively. With *reactive dependent interviewing*, respondents are first asked an independent question, without reference to prior data. The computer script then compares the response with the prior data. If the responses differ (e.g., in the case of categorical variables) or differ beyond a pre-defined threshold (e.g., in the case of continuous variables), the computer script prompts a follow-up question to verify whether the change is true (valid). For example, if reported earnings differ by more than +/–10% from the previous interview, the respondent could be asked:

> *May I please just check?—So your earnings have changed from* <fill: amount a> *to* <fill: amount b> *since we last interviewed you on* <fill: date of interview>?

In addition, the respondent could be asked to clarify the reason for the difference, and this information could later be used for data editing.

With *proactive dependent interviewing*, the previous response is incorporated into the question text. This can be used as a boundary before asking the independent question. For example, respondents may be asked:

> *Last time we interviewed you on* <fill: date of interview>, *you reported receiving* <fill: $amount of unemployment benefits> *each month. Have you continued to receive* <fill: $amount unemployment benefits> *each month since* <fill: date of interview>?

Alternatively, the respondent can be asked to confirm the prior information before being asked about the current situation. For example:

> *According to our records, when we last interviewed you on* <fill: date of interview>, *you were* <fill: employment status>. *Is that correct*?

The prior information can also be used to explicitly ask about change. For example:

> *Last time we interviewed you on* <fill: date of interview>, *you said you were working for* <fill: employer name>. *Are you still working for* <fill: employer name>?

Dependent interviewing is mainly used for factual questions. Respondents generally react positively to interviewers acknowledging information they have provided in earlier waves of interviewing. Cognitive studies suggest that the fact that the interviewer has access to their data does not worry the respondent. However, there are precautions that researchers need to take. For example, confidentiality concerns may arise in surveys that allow proxy reporting. Respondents are not always comfortable with the data they have provided being "fed forward" to a different household member in the future wave of interviewing, were some other member to serve as their proxy. In addition, care also needs to be taken that the wording of reactive dependent interviewing questions that query inconsistent responses do not put respondents off. Finally, the added complexity of the questionnaire script means that implementing dependent interviewing is resource intensive, both in terms of programming and script testing.

Annette Jäckle and Mario Callegaro

See also Coding; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Confidentiality; Interviewer-Related Error; Measurement Error; Panel Survey; Proxy Respondent; Reference Period; Respondent Burden; Respondent-Related Error; Seam Effect; Wave

Further Readings

Hoogendoorn, A. W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics*, *20*, 219–232.

Jäckle, A. (2008). Dependent interviewing: A framework and application to current research. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (chapter 6). Hoboken, NJ: Wiley.

Mathiowetz, N. A., & McGonagle, K. A. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics*, *16*, 401–418.

Pascale, J., & Mayer, T. S. (2004). Exploring confidentiality issues related to dependent interviewing: Preliminary findings. *Journal of Official Statistics*, *20*, 357–377.

Dependent Variable

A dependent variable is a variable that is explained by one or more other variables, which are referred to as "independent variables." The decision to treat a variable as a dependent variable may also imply a claim that an independent variable does not merely predict this variable but also shapes (i.e., causes) the dependent variable. For example, in a survey studying news consumption, exposure to television news could serve as a dependent variable. Other variables, such as demographic characteristics and interest in public affairs, would serve as the independent variables. These independent variables can be used to predict television news exposure and also may be investigated as to whether they also cause one's exposure level.

Researchers often face challenges in establishing causality based on survey data. In causal inference, the dependent variable indicates an outcome or effect, whereas the independent variable is the cause of the outcome or effect. In order to conclude that the dependent variable is caused by the independent variable, the relationship between the two must meet three criteria. First, the two variables must be correlated. That is, a change in one variable must be accompanied by a change in the other. In the case of a positive correlation, one variable increases as the other increases. In the case of a negative correlation, one variable increases as the other decreases. For example, higher levels of education may be associated with lower levels of television news viewing, and if so, there would be a negative correlation between the two variables. If the two variables are not correlated, then there is no causal relationship between them.

Second, the dependent variable must follow the independent variable in the timing of its occurrence. For example, a researcher who seeks to show that one's level of education influences one's level of television news viewing would need to show that changes in the latter occurred *after* changes in the former. In some instances, it is relatively easy to ascertain the temporal order of the variables. For instance, if a researcher investigates the relationship between children's academic performance and their parents' education levels, then he or she may be fairly confident in claiming that the former happened after the latter. In other cases, however, the time order is less clear. For example, it may be difficult to determine the temporal ordering of political knowledge and television news viewing.

Third, the observed correlation between the two variables must be genuine—that is, it cannot be explained by other variables. Even if watching television news is positively associated with political knowledge, the relationship may be spurious, from a causal standpoint, if it can be accounted for by another variable, such as political interest. If the positive correlation between television news viewing and political knowledge is due to the fact that the two variables are both positively related to political interest, then the causal relationship may not be valid, and thus is only one of noncausal covariation.

In establishing a causal relationship between a dependent variable and an independent variable, it is not necessary for the independent variable to be the only cause of the dependent variable. In other words, the independent variable can be one of many factors that influence the dependent variable. For example, education levels may influence the amount of television news one consumes even if many other variables (e.g., interest in politics) also affect news watching.

In survey data, causal relationships between a dependent variable and an independent variable are typically probabilistic rather than deterministic. In other words, the relationship will not necessarily be true for all the cases or even for most cases. For example, if education is found to exert a negative influence on television news viewing, this does not mean that each and every highly educated person watches less television news than each and every less educated person. Thus, finding the cases that violate the relationship does not falsify the causal inference.

Researchers usually face two major challenges while using survey data to establish a causal relationship between the dependent variable and the independent variables(s): (1) ascertaining which variable takes place first, and (2) whether the relationship is genuine. For example, a researcher may find that people who behave aggressively watch more violent television programs but be unable to disentangle the causal direction in the relationship. This is especially likely to be true for analyses using cross-sectional survey data in which the two variables in question are measured at the same time rather than at different points in time and are not measured as part of an experimental design. Moreover, one must rule out other plausible factors that may account for the relationship to ascertain that the observed relationship between the two variables is possibly a causal one. If a nonexperimental survey does not measure all variables that may explain the relationship, then the researcher may not be able to rule out alternative explanations.

Surveys do lend themselves to experimental designs in which the causal relationships between the dependent variable(s) and independent variable(s) can be tested formally. For example, survey researchers can examine experimentally whether response rates are influenced by different levels of incentives or new alternative forms of interviewer persuasion techniques. However, too often survey researchers do not deploy such experimental designs, thus missing the opportunity to better understand the dependent variable(s).

Xiaoxia Cao

See also Experimental Design; Independent Variable; Internal Validity; Noncausal Covariation; Variable

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Kenski, K. (2006). Research design concepts for understanding the rolling cross-section approach. In D. Romer, K. Kenski, K. Winneg, C. Adasiewicz, & K. H. Jamieson (Eds.), *Capturing campaign dynamics 2000 and 2004* (pp. 43–67). Philadelphia: University of Pennsylvania Press.

Schutt, R. K. (2006). *Investigating the social world: The process and practice of research* (5th ed.). Thousand Oaks, CA: Pine Forge Press.

DESIGNATED RESPONDENT

Designated respondents are the individuals chosen specifically to be interviewed for a survey. Surveys often are conducted in two stages: first, selecting a sample of household units and, second, selecting persons within the households with whom to speak. Survey researchers' and interviewers' jobs would be easier if they could question the persons first answering the phone or first coming to the door or simply any adult resident in the unit who was willing to talk. This usually is an acceptable idea only if the researchers simply need to know the basic characteristics of the household; however, much of the time researchers need to gather data from one specifically chosen person in the household—that is, translate the sample of units into a sample of individuals. In contrast, if the respondent is merely the most likely person to answer the phone or to be home, his or her characteristics may be overrepresented in the sample, meaning that the sample will be biased. These more willing or available individuals tend to be older and/or female. Such biases mean that survey researchers are likely to get an inaccurate picture of their samples and can come to some incorrect conclusions. Information quality depends on who is providing it.

Researchers try to avoid such bias by using a within-household selection procedure likely to produce a more representative sample at the person level. These tend to be more expensive than interviewing any available person in the household, but they are also more precise. It takes more time to find the "right person" and to gain an interview when that person is available. As a result, refusal rates can, and often do, increase. The informant (person who answers the door or phone) may be put off by some of the questions interviewers have to ask in order to pick the designated respondent—for example, a complete list of household residents—and may refuse to proceed further. If informants are cooperative but are not the designated respondent, a handoff must occur, and interviewers may have to keep calling back if the designated respondent is not immediately available. Survey researchers have to make trade-offs when they choose a respondent selection method. Different kinds of respondent selection methods have been devised to identify the correct person for interviewing and obtain his or her cooperation, and each has advantages and disadvantages with respect to costs and precision.

Respondent designation techniques have consequences for errors of nonresponse, such as not finding the correct person, inability of the person selected to participate because she or he does not qualify (e.g., because of language barriers, ill health, illiteracy), or that person's unwillingness to be interviewed. Ways to compensate for these problems exist, such as callbacks, interviewing a secondary person in the household who also meets appropriate criteria (e.g., speaks English, is able-bodied, literate), or weighting responses by appropriate criteria. Among principal concerns are within-unit coverage errors; for instance, when the wrong types of respondents consistently are interviewed or when the selected respondents consistently do not meet the survey requirements and another qualified person is available but not interviewed. Survey researchers need to think out solutions to these issues in advance.

Many studies have compared two or more different within-unit selection methods to aid researchers in decisions about procedures that will best fit their needs, although more research on these issues is desirable. This is because some methods of respondent selection violate the principle of random sampling but appear to provide age and sex or other demographic distributions that approximate those in the population of interest. In addition, random sampling should best represent the population of interest, but this does not always happen for a number of reasons.

Usually, the least desirable method is no selection; that is, interviewing whoever answers the phone or door, if age 18 or older (usually adults are the population desired). Although the least expensive method, its common age and gender biases hinder generalizing to the larger population. Data are likely to be less accurate if topics are related to the biases.

The Council of American Survey Research Organizations strongly recommends that market research and attitude studies collect information only by designating a respondent scientifically or according to an appropriate function. Randomness is less of a concern when the designated respondent is, for example, the man of the house, the female head of household, the principal shopper, or the health care decision maker. In cases where informants may indicate that more than one household member qualifies, a random method or other predetermined systematic and unbiased technique will be needed to decide among those qualifying. An example of research on this issue, in 1963, found no significant differences among the four designated respondent procedures that were employed to collect data on home owners' alterations and repairs. The four procedures used in the 1963 study were (1) the head of household, (2) the wife of the head, (3) both together, and (4) any adult in the household with knowledge of these costs. Joint interviews were more difficult to obtain simply because one or the other was more likely to be available than both being available at the same time, and interviewing both persons did not produce a fuller picture than interviews with either one. The researchers speculated that allowing interviewers to ask for the adult best-informed about these consumer expenditures might have been preferable.

Cecilie Gaziano

See also Computer-Assisted Telephone Interviewing (CATI); Hagan and Collier Selection Method; Informant; Kish Selection Method; Last-Birthday Selection; Respondent; Troldahl-Carter-Bryant Respondent Selection; Within-Unit Coverage; Within-Unit Selection

Further Readings

Council of American Survey Research Organizations. (1982). *On the definition of response rates: A special report of the CASRO Task Force on Completion Rates.* New York: Author. Retrieved October 18, 2007, from http://www.casro.org/resprates.cfm

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.
Neter, J., & Waksberg, J. (1963). Effects of interviewing designated respondents in a household survey of home owners' expenditures on alterations and repairs. *Applied Statistics*, *12*, 46–60.
Sabin, M. C., & Godley, S. H. (1987). Mental health citizen surveys: A comparison of two within household telephone sampling techniques. *Evaluation and Program Planning*, *10*, 137–141.
Salmon, C. T., & Nichols, J. S. (1983). The next-birthday method of respondent selection. *Public Opinion Quarterly*, *47*, 270–276.
Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research.* Chicago: University of Chicago Press.

Design-Based Estimation

Design-based estimation methods use the sampling distribution that results when the values for the finite population units are considered to be fixed, and the variation of the estimates arises from the fact that statistics are based on a random sample drawn from the population rather than a census of the entire population.

Survey data are collected to estimate population quantities, such as totals, means, or ratios of certain characteristics. Other uses include comparing subpopulations—for example, estimating the average difference between males and females for certain characteristics. In addition to these descriptive quantities, for many surveys the data are used to fit statistical models, such as linear regression models, to explain relationships among variables of interest for the particular population. In any case, statistics derived from the sample are used to estimate these population quantities, or *parameters*. The basis for assessing the statistical properties of such estimates is the sampling distribution (the probability distribution) of the estimates—the distribution of the estimates that would arise under hypothetical repetitions using the same randomization assumptions and the same form of the estimate.

In design-based estimation, the probabilities used to select the sample are then used as the basis for statistical inference, and such inference refers back to the finite population from which the random sample was selected. These selection probabilities are derived using the particular survey sampling design (e.g., multi-stage, clustered, stratified). In design-based estimation methods, sampling weights are used to account for the possibly unequal probabilities of selection used to draw the sample.

Survey practitioners can also make use of alternative estimation methods including model-based approaches. Pure *model-based estimation* methods assume that the values for the finite population, $Y_1, Y_2, \ldots, Y_N$, are the realization of a random variable from a statistical model, and that the observed outcomes, $y_1, y_2, \ldots, y_n$, can be thought of as having been generated from either that same statistical model or from a statistical model that has been modified to take into account how the sample design has affected the sampling distribution for the sample data. The observations from the sample are used to predict the unobserved units in the population. In contrast, in design-based estimation methods, the values for the finite population units, $Y_1, Y_2, \ldots, Y_N$, are treated as fixed but unknown quantities, and the sampling distribution for the observed outcomes, $y_1, y_2, \ldots, y_n$, arises from the probabilities used to select the units for inclusion into the sample.

Another framework can be used that combines the model and design-based estimation methods and is referred to as a "model-design-based framework" or a "combined distribution." Within this framework, the values for the finite population, $Y_1, Y_2, \ldots, Y_N$, are considered to be the realization of a random variable from a statistical model, and the probability distribution for the outcomes, $y_1, y_2, \ldots, y_n$, is determined by both the statistical model for the population values and the probabilities used to select the units in the sample. Under the model-design-based framework, fitting statistical models to data obtained through a complex survey design, using design-based estimation methods, will often give protection against violation of the model assumptions and any misspecification that may be made with respect to the sampling distribution of the observed data, especially for large sample sizes and small sampling fractions.

Survey-Weighted Estimates

One common outcome in design-based methods is the generation of point estimates that serve to estimate the finite population parameters of interest, such as a population mean, total, proportion, and so on. Such estimates are derived using the sampling weights that are computed in part from the sampling design itself. A simple example to consider here would be the case of selecting a random sample with unequal probabilities of selection from a finite population, where there are no nonresponse and no response errors. In this case, the survey population consists of all units in the population that were eligible for selection in the sample survey design. One assumes that the target population is the same as the survey population. For each unit in the sample, a sampling weight is constructed based on the sampling design. Including this weight for each unit allows one to account for the unequal selection probabilities. When, for each unit in the sample, this weight is equal to the reciprocal of the probability that the unit is included in the sample, the survey-weighted estimate will provide an *unbiased* estimate of the population total. For *multi-stage sampling* designs, the sampling weight is constructed to account for the probabilities of selection at each stage of sampling. An informal interpretation of these weights is that, for each respondent, the weight is approximately equal to the number of units in the population represented by the respondent.

For example, to estimate the population total, $Y = \sum_{i=1}^{N} Y_i$, one could use the survey-weighted estimate given by the statistic $\hat{Y} = \sum_{i=1}^{n} w_i y_i$, where the w_i's are the sampling weights for the observed units. The estimate of the *variance* of this statistic will be based on the design-based sampling distribution of the observations. Statistical inference for large samples (or a large number of primary sampling units in the case of a multi-stage survey design) can be obtained by using the design-based estimate and its estimated design-based variance in conjunction with the normal distribution as an approximation to the sampling distribution of the estimated total. This normal approximation would be the basis for estimating confidence intervals or for conducting statistical hypothesis testing.

The finite population quantities of interest may be more complex than a population total. For example, when the population size is not known, the estimate of a population mean would be the ratio of the survey-weighted estimate of the population total and the survey-weighted estimate of the population size. In this case, the estimate of the population mean would be approximately unbiased. Since the bias tends to zero for large sample sizes, the estimate is said to be *asymptotically design unbiased.* Asymptotically unbiased estimates will be close to their quantities of interest for large samples. Estimates for subpopulations or domains are handled by setting to zero the observed values for all units that fall outside of the domain. Common quantities of interest are domain means or differences between the means of two domains, such as the average difference between males and females for some characteristics of interest.

In practice, there is usually nonresponse, and there may be deficiencies in the sampling frame, such as undercoverage or overcoverage. To account for these deficiencies, adjustments or calibrations are often made to the survey weights. The guiding principle behind such adjustments are to ensure that the survey-weighted estimates are approximately unbiased for the population totals, and possibly to reduce the variance of the estimates. One such example involves using auxiliary data, such as age-sex distributions for the population, to improve the accuracy of the estimates through post-stratification, ratio, or regression estimation.

Analytical Quantities of Interest

When the population of inference is finite, the population quantities of interest are descriptive. However, when fitting a statistical model to survey data, the population of inference is often conceptually infinite, although the population from which samples are drawn are finite. The population of inference is represented by a statistical model from which the values for the finite population units have been generated. The population of inference is larger than the population targeted by the researcher. The quantities of interest are related to the statistical model assumed to have generated the population targeted by the survey taker. In this case, the quantities of interest are analytic, not descriptive.

Design-based estimates for many statistical models are asymptotically design unbiased for the finite population quantities of interest that are associated with the statistical model based on a completely observed finite population. These finite population quantities of interest are usually approximately model unbiased for the parameters of the statistical model. Therefore, the design-based estimates are consistent for the model parameters of interest under the combined or model-design-based framework. The model-design-based variance for the design-based estimate of the model parameter will be close to the design-based variance when the sampling fractions are small and the sample size is large. Therefore, design-based inference for the model parameters of interest would also be valid in the model-design-based or combined framework. Modifications to the design-based variance would be required for cases where the sampling fractions are not negligible.

There are some statistical models for which design-based estimation will not be consistent under the model-design-based framework. These include estimates of the variance components associated with random effects models, mixed effects models, structural equation models, and multi-level models. The fixed effects in these models can usually be estimated consistently, but not the variance components associated with the random effects, unless certain conditions on the sample sizes apply. However, for most models, such as generalized linear models (including linear regression and logistic regression) and proportional hazards models, the parameters of interest can be estimated consistently.

Informative Sampling

The issue of whether a pure model-based estimation approach, as opposed to a design-based estimation approach, is appropriate when estimating quantities from a sample that has been obtained from a complex design is related to whether or not the sampling design is informative. If the sampling distribution of the observations is the same under the model-based randomization assumptions as the sampling distribution under the model-design-based (or combined) randomization assumptions, then the sampling is noninformative. Stratification and clustering in the sample design can lead to informative samples.

When the sampling is informative, the observed outcomes may be correlated with design variables not included in the model, so that model-based estimates of the model parameters can be severely biased, thus leading possibly to false inferences. On the other hand, if the sampling is noninformative, and a design-based estimation approach is used, then the variances of the estimates will usually be larger than the variances of the estimates using a model-based approach.

David A. Binder

See also Auxiliary Variable; Finite Population; Model-Based Estimation; Overcoverage; Parameter; Point Estimate; Population of Inference; Post-Stratification; Probability of Selection; Target Population; Unbiased Statistic; Undercoverage; Variance Estimation; Weighting

Further Readings

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data.* Chichester, UK: Wiley.
Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, *18*(2), 129–154.
Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys.* New York: Wiley.
Särndal, C.-E., Swensson, B., & Wretman, J. (1997). *Model assisted survey sampling.* New York: Springer-Verlag.
Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys.* New York: Wiley.
Smith, T. M. F. (1994). Sample surveys 1975–1990: An age of reconciliation? (with discussion). *International Statistical Review*, *62*(1), 5–34.

Design Effects (*deff*)

The design effect (*deff*) is a survey statistic computed as the quotient of the variability in the parameter estimate of interest resulting from the sampling design and the variability in the estimate that would be obtained from a simple random sample of the same size.

In large-scale sample surveys, inferences are usually based on the standard randomization principle of survey sampling. Under such an approach, the responses are treated as fixed, and the randomness is assumed to come solely from the probability mechanism that generates the sample. For example, in simple random sampling without replacement, the sample mean is unbiased with randomization-based variance given by

$$V_{SRS}(\bar{y}) = (1 - f)\frac{S^2}{n},$$

where n, N, and $f = n/N$ denote the sample size, the population size, and the sampling fraction, respectively, and S^2 is the finite population variance with the divisor $N - 1$. Usually f is negligible and can be dropped from the formula. In any such case, the equality displayed provides a conservative formula for the variance.

In most cases, however, complex sampling designs (indicated by the subscript *CSD* in the following) are applied rather than simple random sampling. In such a situation, $\bar{y}$ can still be an unbiased estimator under the usual randomization approach if the sampling design is one in which each sampling unit in the finite population has the same chance f of being selected. However, $V_{SRS}(\bar{y})$ usually underestimates the true randomization variance of $\bar{y}$ under the complex sampling design, say $V_{CSD}(\bar{y})$. To account for this underestimation, Leslie Kish proposed the following variance inflation factor, commonly known as the design effect:

$$DEFF_R = \frac{V_{CSD}(\bar{y})}{V_{SRS}(\bar{y})}, \tag{1}$$

where subscript R denotes the perspective of the randomization framework. Although in the vast majority of empirical applications, the design effect is considered for the usual sample mean, the ratio in Equation 1 can be defined more generally for the variances of any estimator, θ, under any complex design. In practice, $DEFF_R$ is unknown, and some approximations and estimations are employed to assess its magnitude.

To give an example, consider a population of $N = 9$ elements from which one wishes to select $n = 3$ into the sample. Let the y_i, $i = 1, \ldots, 9$, values be given by 10, 18, 32, 11, 21, 33, 12, 21, 31. If one samples the elements using systematic sampling, as an instance of a complex sample design, exactly three samples are possible: $s_1 = \{10, 11, 12\}$, $s_2 = \{18, 21, 21\}$, $s_3 = \{32, 33, 31\}$. Given these extreme data, it can already be seen, without doing any calculations, that the variance of the sample mean is inflated compared to a simple random sample of three elements. If one calculates the variance of the sample mean given the systematic sample design ($CSD = SYS$), one gets

$$V_{SYS}(\bar{y}) = 74. \tag{2}$$

And, for the variance of the sample mean under simple random sampling,

$$V_{SRS}(\bar{y}) = = \left(1 - \frac{3}{9}\right) \times \frac{84.5}{3} \approx 18.78. \tag{3}$$

Thus the design effect of this example is

$$DEFF_R = \frac{74}{18.78} \approx 3.94, \tag{4}$$

which means that the variance of the sample mean, when choosing the sample by systematic sampling, is nearly 4 times as large as the variance of the same estimator under simple random sampling. This indicates a considerable loss of precision (i.e., larger variance for the same sample size).

It must be noted that the magnitude of the design effect depends on the y values. A different ordering of the values of the study variable in this example yields a different design effect. Now consider the y_i values in the following order: 11, 12, 10, 21, 21, 18, 31, 33, 32, and the possible systematic samples of size 3: $s_1 = \{11, 21, 31\}$, $s_2 = \{12, 21, 33\}$, $s_3 = \{10, 18, 32\}$. Under this ordering of the study variable, the variance of the sample mean given the systematic sample design is

$$V_{SYS}(\bar{y}) = \frac{2}{3} \approx 0.6667,$$

and thus the design effect for the reordered data is

$$DEFF_R = \frac{\frac{2}{3}}{\frac{169}{9}} \approx 0.0355,$$

which implies that in this case systematic sampling is more efficient than simple random sampling (i.e., *design effect* < 1). The reason for these enormous differences lies in the relative homogeneity of the y values within and between the samples.

Systematic sampling is a special case of cluster sampling. The design effect for cluster sampling of n clusters from a population of C clusters, each of size M, can be computed as

$$DEFF_R = \frac{C \cdot M - 1}{M(C - 1)}[1 + (M - 1)\rho], \tag{5}$$

where ρ is the well-known intraclass correlation coefficient (ρ), which is defined as

$$\rho = \frac{\sum_{c=1}^{C} \sum_{j=1}^{M} \sum_{\substack{i=1 \\ j \neq i}}^{M} (y_{ci} - \bar{Y}_c)(y_{cj} - \bar{Y}_c)}{(M - 1)(C \cdot M - 1)S^2}, \tag{6}$$

where y_{cj} denotes the y value of the jth unit in cluster c in the population, and $\bar{Y}_c$ their mean in cluster c. S^2 is the finite population variance of the $C \cdot M$ y values.

The intraclass correlation coefficient can be interpreted as a measure of homogeneity. It ranges from $-\frac{1}{M-1}$ to 1. High values of ρ indicate more homogeneity of y values within the clusters, whereas a low value of ρ indicates less homogeneity. Moreover, negative values indicate a gain in efficiency of the complex design compared to simple random sampling. However, in most empirical applications, ρ takes on small to intermediate values (0.02 to 0.20) depending on the variable under study. In the previous examples, ρ would be computed as 0.978 and –0.487, respectively. Using these values in Equation 5 along with $C = 3$, $n = 1$, and $M = 3$ yields the design effects computed for the original and reordering of the set of 9 y-values, respectively.

In general, design effects that exceed 1 imply less precision per sampled unit for the complex sampling design relative to a simple random sample of the same size, while design effects that are less than 1 imply a gain in precision per sampled unit.

Use of Design Effects

There are several potential uses of design effects. First, design effects are routinely used for the determination of the sample size of a complex survey from knowledge of sample size requirement for a simple random sample design of equal precision. This approach is followed in the European Social Survey (ESS), as described by Peter Lynn, Sabine Häder, Siegfried Gabler, and Seppo Laaksonen. In this context, an important quantity that can be derived from $DEFF_R$ is the effective sample size, n_{eff}, which is defined as

$$n_{eff} = \frac{n}{DEFF_R}. \tag{7}$$

It denotes the corresponding sample size of a simple random sample (more precisely a simple random sample with replacement) that has the same variance as the complex sample design. Usually, n_{eff} is smaller than n, which indicates a loss in precision caused by the complex design. When an overall design effect is known, n_{eff} can be used to compute the sample size, n, of a complex sample, which is required to ensure a pre-defined precision.

In the absence of any direct survey data on the response variables, historical data as well as information from similar surveys are used in conjunction with the information available on the survey under consideration such as average cluster size, number of clusters, and so on.

The second possible use of design effects is for variance computation from complex surveys in situations in which standard variance estimation techniques cannot be employed—either due to unavailability of appropriate software, especially in developing countries, or due to unavailability of actual cluster identifiers to protect the confidentiality of survey respondents. For this use, survey researchers and practitioners often publish design effects of core items together with survey data.

Estimation of Design Effects

In practice, the design effect depends on unknown population quantities that have to be estimated from sample data. Hence, the numerator and denominator of the right-hand side of Equation 1 have to be estimated from the sample data. Estimating the numerator leads to the classical variance estimation problem. In the case of stratified random sampling or cluster sampling, adequate variance estimators are available. However, in complex surveys with unequal probability sampling, second-order inclusion probabilities π_{ij} have to be available. Since the computation of the π_{ij} may be extremely cumbersome, adequate approximations may have to be considered.

The generalization of Equation 1 to calibration estimators or nonlinear statistics generally leads to applying residual or linearization techniques, as discussed by A. Demnati and J. N. K. Rao and by J. C. Deville. Alternatively, resampling methods, such as the jackknife or bootstrap, can be applied in order to build the sampling distributions via estimating from adequate subsamples from the original sample.

The estimation of the denominator of Equation 1 leads to estimating the variance of the given estimator under simple random sampling with the given sample. However, this sample was drawn using a complex sampling design and cannot be directly used for variance estimation under simple random sampling. One way to compensate for unequal probabilities is to estimate S^2 by

$$\hat{S}^2 = \frac{1}{\sum\limits_{i \in S} \pi_i^{-1} - 1} \sum_{i \in S} \frac{1}{\pi_j} \left(y_i - \frac{1}{\sum\limits_{i \in S} \pi_i^{-1}} \sum_{j \in S} \frac{y_j}{\pi_j} \right)^2. \tag{8}$$

Alternatively, one may wish to estimate the population distribution and from this an estimator of S^2.

Model-Based Approach to Design Effects

Model-based estimation differs from the design-based approach mainly in the assumptions about the data-generating process and hence the way estimators of population parameters have to be considered. This approach is mainly helpful in the design stage of a sample survey when no data are available.

A model-based version of the design effect has been suggested by Gabler, Häder, and Parthasarathi Lahiri. Let b_c be the number of observations in the cth of C clusters. Hence, $\bar{b} = \frac{1}{C}\sum_{c=1}^{C} b_c$ is the average cluster size. Taking into account the usual design-based estimator of the population mean, $\bar{y}_w = \frac{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}y_{cj}}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}}$, let us assume the following model (M1):

$$Var_{M1}\left(y_{cj}\right) = \sigma^2 \text{ for } c = 1, \ldots, C;\ j = 1, \ldots, b_c \quad (9)$$

$$Cov_{M1}\left(y_{cj}, y_{c';j'}\right) = \begin{cases} \rho\sigma^2 & \text{if } c = c'; j \neq j'; \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

A second model (M2) specifies the distribution of the y_{cj} in the following way:

$$Var_{M2}\left(y_{cj}\right) = \sigma^2 \text{ for } c = 1, \ldots, C; j = 1, \ldots, b_c \quad (11)$$

$$Cov_{M2}\left(y_{cj}, y_{c';j'}\right) = 0 \text{ for all } (c, j) \neq (c', j'). \quad (12)$$

Let $Var_{M1}(\bar{y}_w)$ be the variance of the weighted sample mean under model M1 and let $Var_{M2}(\bar{y})$ be the variance of the overall sample mean, $\bar{y} = \frac{\sum_{c=1}^{C}\sum_{j=1}^{b_c} y_{cj}}{C\bar{b}}$, under M2. Under M2, the variance of $\bar{y}$, however, turns out to be given by $Var_{M2}(\bar{y}) = \frac{\sigma^2}{n}$. Then the model-based design effect is defined as

$$DEFF_M = \frac{Var_{M1}(\bar{y}_w)}{Var_{M2}(\bar{y})}. \quad (13)$$

After some algebra, it turns out that $DEFF_M$ can be expressed as

$$DEFF_M = n\frac{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2}{\left(\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}\right)^2} \times [1 + (b* - 1)\rho], \quad (14)$$

where

$$b* = \frac{\sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2}. \quad (15)$$

The first term of Equation 14 is the design effect due to unequal selection probabilities, $DEFF_P$, and the second term is the well-known design effect due to clustering, $DEFF_C$. Thus, Equation 1 can equivalently be written as a product of $DEFF_P$ and $DEFF_C$:

$$DEFF_M = DEFF_P \times DEFF_C. \quad (16)$$

The quantity ρ again servers as a measure of homogeneity. The usual analysis of variance (ANOVA) estimator of $\rho||\text{I}||$ is given by

$$\hat{\rho}_{ANOVA} = \frac{MSB - MSW}{MSB + (K - 1)MSW}, \quad (17)$$

where

$$MSB = \frac{SSB}{C - 1}$$

with $SSB = \sum_{c=1}^{C} b_c(\bar{y}_c - \bar{y})^2$, $\bar{y}_c$ the sample mean of the y values in the cth cluster, and

$$MSW = \frac{SSW}{n - C}$$

with $SSW = \sum_{c=1}^{C}\sum_{j=1}^{b_c}(y_{cj} - \bar{y}_c)^2$ and

$$K = \frac{n - \frac{\sum_{c=1}^{C} b_c^2}{n}}{C - 1}.$$

In simulation studies, the ANOVA estimator is usually found to be an approximately unbiased, efficient, and consistent estimator of ρ, as discussed by S. R. Paul, K. K. Saha, and U. Balasooriya. These empirical findings, together with its appealing intuitive interpretation and its computational simplicity, are the reasons why it is used in the estimation of design effects in many surveys (e.g., the ESS).

The model described has the advantage that it applies to many real-world situations. In the ESS, for example, the model-based design effect is estimated according to the above formula in countries where sampling was done using (a) unequal selection probabilities, (b) clustering, or (c) both. What makes it even more useful is that it can also be applied to multiple design samples. Gabler, Häder, and Lynn showed that Equation 1 has a generalized form that allows a weighted average to be calculated over multiple domains in a sample.

Software

Today, most of the popular statistical software packages offer an option for data analyses to allow for complex designs—either by providing an estimate of the design effect or by their capability to account for the complex design in the variance estimation. These include STATA, SUDAAN, and WesVar PC.

Siegfried Gabler, Matthias Ganninger, Sabine Häder, and Ralf Münnich

See also Bootstrapping; Cluster Sample; Complex Sample Surveys; Design-Based Estimation; Effective Sample Size; Jackknife Variance Estimation; Model-Based Estimation; ρ (Rho); Sample Design; Systematic Sampling; Unbiased Statistic; Variance Estimation; WesVar

Further Readings

Cohen, S. B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistician*, *51*(30), 285–292.

Davison, A. C., & Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, *23*(3), 371–386.

Demnati, A., & Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, *30*(1), 17–26.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, *25*(2), 193–203.

Gabler, S., Häder, S., & Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, *25*(1), 105–106.

Gabler, S., Häder, S., & Lynn, P. (2006). Design effects for multiple design surveys. *Survey Methodology*, *32*(1), 115–120.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lynn, P., & Gabler, S. (2005). Approximations to b* in the prediction of design effects due to clustering. *Survey Methodology*, *31*(2), 101–104.

Lynn, P., Häder, S., Gabler, S., & Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: The European Social Survey experience. *Official Statistics*, *23*(1), 107–124.

Paul, S. R., Saha, K. K., & Balasooriya, U. (2003). An empirical investigation of different operation characteristics of several estimators of the intraclass correlation in the analysis of binary data. *Journal of Statistical Computation & Simulation*, *73*(7), 507–523.

Diary

A diary is a type of self-administered questionnaire often used to record frequent or contemporaneous events or experiences. In *diary surveys*, respondents are given the self-administered form and asked to fill in the required information when events occur (event-based diaries) or at specified times or time intervals (time-based diaries). Data from diary studies can be used to make cross-sectional comparisons across people, track an individual over time, or study processes within individuals or families. The main advantages of diary methods are that they allow events to be recorded in their natural setting and, in theory, minimize the delay between the event and the time it is recorded.

Diaries are used in a variety of domains. These include studies of expenditure, nutrition, time use, travel, media exposure, health, and mental health. Expenditure surveys usually have a diary component in which the respondent has to enter expenditures on a daily basis for a short period of time, such as a week or 2 weeks. An example of this is the Consumer Expenditure Survey in the United States, in which one household member is assigned two weekly diaries in which to enter household expenditures. Food and nutrition surveys use diaries to record food consumption over a fixed period of time. An example is the 1996 Food Expenditure Survey in Canada.

Types of Diaries

Time-use diaries usually have shorter reference periods than expenditure diaries. The most common methodology is a diary where the respondent accounts for all his or her activities in a period of 24 hours. If different respondents get assigned different days, the data are used to construct a synthetic week using data from other respondents with similar characteristics. Sometimes, respondents are asked to record their activities at random times during the day when they are signaled by an electronic device. In other time-use surveys, the diary is used as a recall aid for in-person or phone interviews. Time-use researchers have often found that when people are asked about what they spend time on, they often overestimate or underestimate time spent relative to what they actually record in diaries.

Travel surveys use diaries to record trips. Some examples are the 2001 National Household Travel Survey, which recorded information about one travel day, and the 1995 American Travel Survey, which was a 3-month travel survey structured in the form of a calendar. Media exposure diaries are used by companies in the United States like Nielsen and Arbitron to measure the size and composition of the television and radio audiences, respectively, in specific geographic media markets. The Nielsen TV Diary covers television tuning and viewing for all household members in their home for a 7-day week, while the Arbitron radio diary is for one person and covers radio listening anywhere it may take place during a 7-day week.

Diaries are also widely used in health, mental health, and by researchers in various areas of psychology. Diary studies have been used to investigate symptoms, medications, pain levels, substance use, unsafe sexual practices, depression, anxiety, addictions, use of health services, and many other medical issues.

Paper-and-pencil diaries are the oldest kind of diary instrument and can be structured in different ways depending on the type of survey. Paper-and-pencil diaries can be of a journal type (which are unstructured), product type (in categories), outlet type (by place), or day/time type (which covers each hour or minute of each day in the measurement period). An ideal paper-and-pencil diary would be portable, incorporate simple instructions, and have an appropriate level of structure and organization. Though they are very easy to use, paper diaries can be problematic. Respondents often forget to fill them out in a timely manner and later make recall errors. The burden of data entry and processing can be heavy for these diaries. Augmented paper diaries are sometimes used in time-based surveys, when respondents record in a paper diary, and a device like a beeper or pager, programmable wristwatch, or phone call reminds them to fill out the diary.

Advantages and Disadvantages

Recent technological innovations in diary studies include the use of handheld devices, voice activated recorders, scanners, and Web-based diaries. Some devices now in use include handheld computers, personal digital assistants, and electronic diaries. Electronic devices have the benefit of being portable, can have time and date stamps, and are easy to program to allow for signaling or other kinds of customization. Although data entry is easier, the costs of training, program development, hardware, and repairs can be quite high.

There are several problems with diary surveys in general. Since participation often involves a large time commitment, response rates are can be very low. Additionally, there are problems with accuracy of data entry by respondents. Errors include forgetting to fill the diary or filling it in erroneously because of recall problems caused by delay. The process of having to fill out a diary may also affect the respondent's behavior. For instance, respondents may change their levels of food consumption in food surveys or purchase fewer items in expenditure surveys during the time they are participating in the diary survey merely because they know they are being measured. Finally, diary studies can be expensive both because of the cost of the technological devices and also the costs of interviewers having to make repeated visits to train respondents to use the diary, monitor respondents to ensure that they fill it out, and pick it up at the end of the survey.

Parvati Krishnamurty

See also Aided Recall; Questionnaire Design; Respondent Burden; Respondent-Related Error; Survey Costs

Further Readings

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616.

Butcher, R., & Eldridge, J. (1990). The use of diaries in data collection. *The Statistician*, *39*(1), 25–41.

Sudman, S., & Ferber, R. (1971). Experiments in obtaining consumer expenditures by diary methods. *Journal of the American Statistical Association*, *66*, 725–735.

DIFFERENTIAL ATTRITION

Panel studies are subject to *attrition*, which is unit nonresponse after the initial wave of data collection. Attrition affects the results of analyses based on panel data by reducing the sample size and thereby diminishing the efficiency of the estimates. In addition, and more important, attrition also may be selective; differential or selective attrition occurs when the characteristics of the panel members who drop out of the panel because of attrition differ systematically from the characteristics of panel members who are retained in the panel. Differential attrition may introduce bias in survey estimates. However, the amount of bias depends both on the amount of attrition and on the selectivity of attrition, or in other words, on the association between the variables from which the estimate is constructed and the attrition propensity of the panel units. If an estimate is not associated at all with the attrition propensity, then the data are not biased. However, if an estimate is associated with the propensity to participate in the panel, the data are biased.

The propensity to participate in a panel survey (or alternatively, the propensity to be contacted, and given contact, the propensity to agree to participate in the panel survey) is influenced by many different factors, from characteristics of the survey design, survey-taking climate, and neighborhood characteristics to sociodemographic characteristics of the sample persons, the sample persons' knowledge of the survey topic, and their prior wave experiences. For example, the "at-home" patterns of a household and its members, and thus also their propensity to be contacted, are a function of sociodemographic attributes (e.g., number of persons in household) and lifestyle (e.g., working hours, social activities). If one person lives alone in a housing unit, contact is completely dependent on when he or she is at home. Likewise, the lifestyles of younger people may involve more out-of-home activities than those of other groups, and this also means that they will be harder to contact. Consequently, for example, when studying the extent of and changes in social contacts as teenagers grow into adulthood and later when they start their own families, the results are likely to be biased because the survey disproportionally loses (due to attrition) young individuals with more out-of-house activities. A similar logic underlies how error related to refusals is generated. For example, some studies of panel attrition provide evidence that a pleasant survey experience enhances the chance that people will participate in subsequent surveys, whereas those without such an experience are less likely to participate. Participating in a survey is a negative experience when one lacks the cognitive ability to perform the respondent task. We can assume that respondents with low socioeconomic status, including lower educational attainment, might have more difficulties in performing the respondent task; consequently, the interview is an unpleasant or bad experience, and these respondents will be less motivated to participate again in the panel survey. Since socioeconomic status is an important explanatory variable in many panel data analyses, it may be expected that at least some of the conclusions of these studies will be based on biased estimates due to the resulting differential attrition.

Attrition may also be selective with respect to the recent behavior of panel members or recent changes in their position, for example, a divorce transition. Several attrition studies have shown that noncontact is more likely after a household move. However, the move itself is usually precipitated by a particular set of circumstances, and specific events, such as marriage or divorce, affect the likelihood of moving. A divorce is also a stressful situation and can cause a family crisis, which may prevent panel members from participating in a new wave of the panel survey. Since there might be a relationship between the propensity to undergo the change being analyzed, that is, getting a divorce, and the propensity to leave the panel survey, the divorce propensity estimated on the basis of the panel data is most likely an underestimate of the real divorce propensity.

Femke De Keulenaer

See also Attrition; Nonresponse Bias; Panel Survey; Unit Nonresponse; Wave

Further Readings

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources*, *33*(2), 251–299.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (Eds.). (1989). *Panel surveys.* New York: Wiley.
Winkels, J. W., & Davies Withers, S. (2000). Panel attrition. In D. Rose (Ed.), *Researching social and economic change: The uses of household panel studies* (pp. 79–95). New York: Routledge.

Differential Nonresponse

Differential nonresponse refers to survey nonresponse that differs across various groups of interest. For example, for many varied reasons, minority members of the general population, including those who do not speak as their first language the dominant language of the country in which the survey is being conducted, are generally more likely to be nonresponders when sampled for participation in a survey. Thus, their response propensity to cooperate in surveys is lower, on average, than that of whites. The same holds true for the young adult cohort (18–29 years of age) compared to older adults. This holds true in all Western societies where surveys are conducted.

Ultimately, the concern a researcher has about this possible phenomenon should rest on whether there is reason to think that differential nonresponse is related to *differential nonresponse error*. If it is not, then there is less reason for concern. However, since nonresponse error in itself is difficult to measure, differential nonresponse error is even more of a challenge.

In considering what a researcher should do about the possibility of differential nonresponse, a researcher has two primary options. First, there are things to do to try to avoid it. Given that noncontacts and refusals are typically the main causes of survey nonresponse, researchers can give explicit thought to the procedures they use to make contact with respondents (e.g., advance letters) and those they use to try to avoid refusals from respondents (e.g., refusal conversation attempts)—in particular as these procedures apply to key groups from whom lower levels of contact and/or cooperation can be expected. For example, the use of differential incentives to persons or households known from past research to be harder to contact and/or gain cooperation from has been shown to be effective in lowering differential nonresponse. However, some have argued that it is not "equitable" to provide higher incentives to groups that traditionally have low response rates because it fails to fairly "reward" those who readily cooperate in surveys.

However, an unpleasant paradox exists for those who argue that differential strategies aimed at reducing differential nonresponse are inequitable to those respondents who are easier to contact and/or more readily cooperate. When a new treatment (e.g., higher noncontingent incentives) is implemented across the board to raise response rates—so that everyone gets the same treatment—it often increases the gap in response rates between the lowest responding groups and the highest responding groups rather than narrowing the gap between the two groups. This results in an *increase in the size of the differential nonresponse.*

The second option for researchers is to use a variety of post-survey adjustments to their raw data to account for differential nonresponse. If there is no differential nonresponse error associated with the differential nonresponse, then these adjustments will likely be adequate. However, too often it is not known whether there is any error associated with the differential nonresponse, and thus researchers cannot know with confidence whether their adjustments have accomplished anything to help make the survey more accurate.

Paul J. Lavrakas

See also Advance Letter; Nonresponse; Nonresponse Error; Refusal Conversion; Response Propensity

Further Readings

Little, T. C., & Gelman, A. (1998). Modeling differential nonresponse in sample surveys. *Sankhyā: The Indian Journal of Statistics, 60*, 101–126.
Murphy, W., O'Muircheartaigh, C., Harter, R., & Emmons, C. (2003, May). *Optimizing call strategies in RDD: Differential nonresponse bias and costs in REACH 2010.* Paper presented at the 58th Annual Conference of the American Association for Public Opinion Research, Nashville, TN.
Singer, E., Groves, R. M., & Corning, A. (1999). Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly*, *63*, 251–260.
Trussell, N., & Lavrakas, P. J. (2004, May). *Using larger cash incentives to reduce non-response among hard to reach targeted demographic subgroups: It depends on how you pay it.* Paper presented at the 59th Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.
Williams, W. H., & Mallows, C. L. (1970). Systematic biases in panel surveys due to differential nonresponse. *Journal of the American Statistical Association*, *65*(331), 1338–1349.

Directory Sampling

Directory sampling is one of the earliest versions of telephone sampling. Telephone directories consist of listings of telephone numbers. The residential numbers are generally placed in a section of the directory separate from business numbers. Each telephone listing is generally accompanied by a name and an address, although the address is not always present. Households may choose not to have their telephone number published in the directory. These are referred to as *unpublished numbers*, most of which also are *unlisted numbers*.

In the original application of directory sampling, a set of telephone directories covering the geopolitical area of interest to the survey were assembled. After the sample size of telephone numbers was determined, a random selection procedure was used to draw the required number of residential directory-listed telephone numbers for each directory. The actual selection method ranged from using systematic random sampling of listed telephone numbers to first selecting a sample of pages from the directory and then sampling one or more telephone numbers from the selected pages.

Directory samples provide samples only of telephone numbers that are directory listed. Directory samples will yield biased samples of a population, because all unlisted households are given a zero probability of selection, and unlisted households generally differ from listed households on key characteristics. For example, persons with unlisted numbers are more likely to be minorities, recent movers, and single female adults. In some geographic areas, a substantial percentage of households may have unlisted telephone numbers, for example, larger central city areas and Western states.

Today, directory-listed sampling is rarely used alone, having been replaced by list-assisted random-digit dial sampling. But in other ways, directory sampling has made a comeback. Telephone directories are now entered into national databases of listed residential telephone numbers that are updated on an ongoing basis. A fairly common random-digit dialing sample design involves forming two strata. The first stratum consists of directory-listed residential telephone numbers. The second stratum consists of telephone numbers in the list-assisted sampling frame that are not residential directory-listed telephone numbers. Thus two mutually exclusive strata are formed, and a sample of telephone numbers is drawn from each stratum.

The presence of an address for most residential directory listed telephone numbers in national databases makes it possible to assign geographic codes to the addresses. Typical geographic codes include county, zip code, census tract, block group, and census block. This makes it possible to sample directory-listed telephone numbers from small geographic areas, for example, from a *reverse directory*. The presence of a name with each listed number also enables the matching of the names to lists of ethnic surnames. This makes it possible to sample directory-listed households with specific surnames.

Michael P. Battaglia

See also List-Assisted Sampling; Random-Digit Dialing (RDD); Reverse Directory; Systematic Random Sample; Telephone Survey; Unlisted Household; Unpublished Number

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Disclosure

Within the context of survey research, *disclosure* can be used with two distinct meanings. In the first meaning, a researcher is required to provide full disclosure of his or her own identity and purpose in collecting data. In the second meaning, a researcher is required to prevent disclosure of information that could be used to identify respondents, in the absence of specific and explicit informed consent allowing the researcher to disclose such information.

However, in some research settings, full disclosure of the research objectives may jeopardize the objectivity of results or access to research participants. Observational research of behavior in public settings, for example, may be exempt from rules of informed consent, since the public nature of the behavior itself implies consent. Nevertheless, in this situation, the researcher ideally should provide detailed justification for the data collection methodology in any research proposal or data presentation, and the methodology should be subject to peer and ethical review. In addition, the participants' right to privacy, anonymity, and confidentiality gains additional importance in such cases, since respondents have not given explicit consent and are not cognizant of the purpose or objective for which they "provide" information. Whenever possible, participants should be debriefed as to the research objectives and use of data after completion of the research or observation and given the opportunity to refuse participation.

Another situation that challenges many researchers in the effort to fully disclose their role and objective as researchers is one in which gatekeepers are involved. When gatekeepers control access to the participants of the research, full disclosure to the gatekeeper is necessary but not sufficient to gain access to the research participant. Permission obtained from the gatekeeper may not be substituted for the need to take separate and full informed consent of the participants. The rights of participants in such situations are the same as in all other cases and need determined protection.

In the second use of the term, disclosure of a respondent's identity or identifying information is prohibited in the absence of specific, informed consent. Research for which disclosure of the subject's identity and/or responses could put the individual at risk of criminal or civil liability or damage the subject's financial standing, employability, or reputation is especially problematic and is generally subject to review by an institutional review board.

Disclosure risks may involve a direct risk, when the disclosure of a respondent's identity or responses may cause harm to the respondent because of the nature of the data themselves, or the risk may be indirect, when risk involves the potential for combining the collected data with an external database through which individuals may be identified and confidential information exposed. This indirect disclosure risk is becoming far more problematic nowadays with the availability of many various data sources, and respondent protections are increasingly focused on this second type of disclosure risk.

Recent expansion in the aggregation of data from a variety of sources that link individuals using identifying information has increased researchers' concerns about confidentiality protection and the disclosure of research subjects' identity. Although confidentiality is promised in the data collection process, the obligations of those disseminating "cleaned" data sets are often less formal and less clear. As commercial databases that include names, addresses, and other sensitive information have become more accessible, the potential for misuse has grown.

When data sets are made public or disseminated, any codes or variables that can be used in combination to isolate and identify a small population subgroup or class pose a risk of disclosure. Ethnicity, for example, in combination with age, gender, and a detailed occupational group or specific geographic identifier may provide sufficient information to disclose an individual identity. Some protections to reduce the likelihood that this form of disclosure may occur include the following:

1. *Coarsening* the data set involves disguising identifying information within a data set. Variables such as age may be rounded in order to remove the precision that might allow for identification. Income is a visible and highly sensitive characteristic that may be top and bottom coded, so that each income extreme, whether for households, persons, or families, including total income and its individual components, is combined into "over" and "under" categories.

2. *Microaggregation* is the process of creating artificial respondents synthesized from averaged responses. For the Substance Abuse and Mental Health Services Administration's Alcohol and Drug Services Study (ADSS) groups, cases in sets of three for problematic variables could potentially be linked to other files or could be used to identify an individual or organization. The average of the three records for each grouping is then recorded as the record for each case in the group.

3. *Suppression* is the removal of any estimate or value in which cells are below a certain size. For example, the Census Bureau and National Center for Health Statistics require that all geographic areas identified must have at least 100,000 persons in the sampled area (according to latest census or census estimate). Other variables, such as duration of residence, migration specifying movement from one type of area to another, distance of a residence from an identified geographic area, or the existence of a particular service or utility (such as well water, septic tanks, and cable TV) for which only a small area has or does not have this type of service are also treated as sensitive variables capable of disclosing respondent

identity and suppressed from publicly disseminated data files.

Laws generally do not protect researchers from disclosure in the ways that journalist–sources, lawyer–client communications, and doctor–patient relationships are often exempted from required disclosures of identify and content of communication. Researchers are ethically required actively to protect respondents' identities, particularly when data sets may be distributed, combined, or used in other, unforeseen ways.

Amy Flowers

See also Confidentiality; Cell Suppression; Informed Consent; Privacy

Disclosure Limitation

Survey researchers in both the public and private sectors are required by strong legal and ethical considerations to protect the privacy of individuals and establishments who provide them with identifiable information. When researchers publish or share this information, they employ statistical techniques to ensure that the risk of disclosing confidential information is negligible. These techniques are often referred to as "disclosure limitation" or "disclosure avoidance" techniques, and they have been developed and implemented by various organizations for more than 40 years.

The choice of disclosure limitation methods depends on the nature of the data product planned for release. There are specific disclosure limitation methods for data released as micro-data files, frequency (count) tables, or magnitude (point estimates) tables. Online query systems may require additional disclosure limitation techniques, depending on whether the data underlying these systems are in the form of micro-data files or tables.

The first step in limiting disclosures in data products is to delete or remove from the data any personal or "direct" identifiers, such as name, street address, telephone number, or Social Security number. Once this is done, statistical disclosure limitation methods are then applied to further reduce or limit disclosure risks.

After direct identifiers are deleted from a micro-data file, there is still a possibility that the data themselves could lead to a disclosure of the individual, household, or business that provided them. Some people and some businesses have unique characteristics that would make them stand out from others. Applying micro-data disclosure limitation methods reduces the possibility of locating these unique records. Some of these methods are data reduction (delete data fields or records), data swapping, micro-aggregation, data perturbation, and imputation.

Protected micro-data produce protected tables. However, sometimes there is interest in producing tables without changing the underlying micro-data. Disclosure limitation methods for tables are applied directly to the tables. These methods include redesign of tables (collapsing rows or columns), cell suppression, controlled and random rounding, and synthetic data substitution.

The application of most disclosure limitation methods will result in some loss of information. Survey researchers should carefully select the appropriate disclosure limitation methods not only to maximize the information retained and the benefits accrued through data release but also protect confidential information from disclosure. However, when judging the risks of disclosure against the loss of information and the benefits of data release, survey researchers should recognize that there is no way to ensure complete elimination of disclosure risk short of not releasing any tables or micro-data files.

Stephen J. Blumberg

See also Cell Suppression; Confidentiality; Data Swapping; Imputation; Perturbation Methods

Further Readings

Federal Committee on Statistical Methodology. (2005). *Statistical policy working paper 22 (Second version): Report on statistical disclosure limitation methodology.* Washington, DC: Office of Management and Budget. Retrieved March 29, 2008, from http://www.fcsm.gov/working-papers/spwp22.html

Disk by Mail

Disk by mail is a survey administration technique in which a selected respondent is mailed a computer disk that contains a questionnaire and a self-starting interview program. The respondent runs the program on his or her own computer and returns the disk containing the completed questionnaire. In some instances, the disk may provide an option for the person to transmit his or her responses over the Internet. Although disk-by-mail surveys can be conducted with the general public, the approach is most effective for targeted populations such as professional or business groups for whom computer access is nearly universal.

Disk by mail is one of a variety of computer-assisted self-interview (CASI) techniques. As such it has some of the advantages of a computerized survey. These surveys have the capability of guiding the respondent interactively through the questionnaire and including very complex skip patterns or rotation logic. This approach can also offer many innovative features beyond traditional mail and telephone surveys, but it does require costs and time in terms of programming and distribution of the survey. Because the approach is computer based, it allows the researcher to enhance the survey forms with respect to the use of color, innovative screen designs, question formatting, and other features not available with paper questionnaires. They can prohibit multiple or blank responses by not allowing the participant to continue on or to submit the survey without first correcting the response error.

Disk by mail also shares some of the advantages of mail surveys. It is less expensive than telephone surveys since there are no interviewer costs incurred, eliminates the potential for interviewer bias, provides respondents with greater "perceived" anonymity that may lead to more truthful answers, especially on sensitive questions; and allows respondents to complete the survey on their own time, that is, when it is most convenient.

Disk by mail does have some drawbacks as a survey technique. It is restricted to those having access to a computer and limited by the technological capacity or make of the respondent's computer. Although disk-by-mail surveys allow for much more innovative features than paper-and-pencil mailed surveys, some respondents may have difficulty accessing the survey due to poor computer skills and will not be able to respond. Furthermore, some people are not accustomed to the process used to respond to an electronic survey (e.g., selecting from a pull-down menu, clicking a radio button, scrolling from screen to screen) and will need specific instructions that guide them through each question and the manner in which they should respond. As with other computer-based survey tools, respondents are often concerned about confidentiality and may be reluctant to download files in fear that they may contain viruses. Additionally, disk by mail typically requires a longer fielding period than some other methods (such as telephone) to complete the project, can make it difficult for the respondent to ask questions or seek clarification, can be limited by low literacy rates among some populations, and provides researchers with little control over who actually completes the survey, thus leading to the possibility of within-unit coverage error.

Michael W. Link

See also Anonymity; Coverage Error; Computer-Assisted Self-Interviewing (CASI); Confidentiality; Radio Buttons; Within-Unit Coverage Error

Further Readings

Couper, M. P., & Nichols, W. L. (1998). The history and development of computer assisted survey information collection methods. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. E. Clark, J. Martin, W. L. Nichols, et al. (Eds.), *Computer assisted survey information collection* (pp. 1–22). New York: Wiley.

De Leeuw, E., Hox, J., & Kef, S. (2003). Computer-assisted self-interviewing tailored for special populations and topics. *Field Methods*, *15*, 223–251.

Saltzman, A. (1993). Improving response rates in disk-by-mail surveys. *Marketing Research*, *5*, 32–39.

Dispositions

Sample dispositions (codes or categories used by survey researchers to track the outcome of contact attempts on individual cases in the sample) provide survey researchers with the status of each unit or case within the sampling pool and are an important quality assurance component in a survey, regardless of the mode in which the survey is conducted. Sample dispositions are used for three reasons: (1) to help the survey researcher control the sampling pool during the field period, (2) to calculate response rates, and (3) to help assess whether the sample might contain nonresponse error. Sample dispositions usually are tracked through the use of an extensive system of numeric codes or categories that are assigned to each unit in the sampling pool once the field period of the survey has begun. Common sample dispositions include the following:

- Busy (telephone survey)
- Fast busy (telephone survey)
- Callback
- Completed interview
- Household refusal
- Ineligible respondent
- Ineligible household
- Language barrier
- Nonresidential address or number (in-person, mail, and telephone surveys)
- Nonworking number (telephone survey)
- Partial interview
- Privacy manager (telephone survey)
- Respondent refusal
- Ring–no answer (telephone survey)
- Unable to participate
- Unavailable respondent
- Unpublished number (telephone survey)
- Vacant housing unit (mail and in-person surveys)

Sample dispositions usually are assigned each time a case is contacted during the field period of a survey, and final sample dispositions are assigned once the field period of a survey has been completed (and the final status of each case in the sampling pool is known). For example, the disposition code of each telephone number in the sample for a telephone survey is updated after every call that is made to the number by an interviewer. In the case of a mail survey, sample dispositions may be updated as completed survey questionnaires are returned to researchers by respondents or as the postal service brings mail "returned to sender" back to the researchers in the case of incorrect addresses or respondents who have moved. In an Internet survey, sample dispositions may be updated as email invitations are sent to individuals in the sampling pool, as email nondelivery messages are returned to the sender after not being able to be delivered (in the case of an incorrect or nonworking email address), as respondents log in to complete the Web survey, and as respondents complete the questionnaires. In an in-person household survey, sample dispositions may be updated as field interviewers visit addresses listed in the sampling frame.

Uses of Sample Dispositions

Controlling the Sampling Pool

A primary purpose of sample dispositions is to assist researchers in controlling the sampling pool during the field period for a survey. For example, if interviewers in an in-person survey were allowed to choose neighborhoods and households to visit from the sampling pool, nonresponse error would be likely because the resulting sample would include a disproportionately large number of households in neighborhoods that are easy to reach, consist primarily of single-family houses, and have higher socioeconomic statuses (and thus may be safer to visit). Sample dispositions make it possible for survey managers to ensure that all cases in the sampling pool are contacted at a variety of days and times and that specific appointments are kept. In this way, sample dispositions help researchers maximize the efficiency of interviewers. In telephone, mail, Internet, and in-person surveys, sample dispositions make it possible for survey managers to ensure that nonresponding cases in the sampling pool are targeted by follow-up mailings, reminder emails, telephone calls, or visits by interviewers.

Calculating Response Rates

Another important purpose of sample dispositions is to calculate survey response rates. It is standard practice for survey researchers to compute a survey's response rates at the end of the field period. Response rates are a common measure of survey quality, and it often is assumed that the higher the response rate, the higher the quality of the survey data. Because sample dispositions categorize the outcome of each case (or unit) in the sampling pool, sample dispositions make it possible for survey researchers to calculate survey response rates.

Assessing Nonresponse Error in the Sampling Pool

A third important purpose of sample dispositions is to assess potential nonresponse error in the sampling pool. Correct or not, a common assumption is that there is more nonresponse error in survey samples with lower response rates than in survey samples with higher response rates. Although determining the amount of nonresponse error in survey data requires more than just knowing the survey response rate, calculating survey response rates is an important first step in understanding whether nonresponse error is present in survey data.

Matthew Courser

See also Field Period; Final Dispositions; Nonresponse Error; Paradata; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Disproportionate Allocation to Strata

One type of random sampling employed in survey research is the use of disproportionate allocation to strata. Disproportionate allocation to strata sampling involves dividing the population of interest into mutually exclusive and exhaustive strata and selecting elements (e.g., households or persons) from *each* stratum.

Commonly used strata include geographic units; for example, high-minority-density census tracts in a city are put into one stratum and low-minority-density census tracts are put into another stratum. In epidemiology case-control studies, strata are used where persons in one stratum have a condition of interest (e.g., Type I diabetes) and persons without the condition are put into a second stratum. After dividing the population into two or more strata, a "disproportionate" number of persons are selected from one stratum relative to others. In other words, the persons in one stratum have a higher probability of being included in the sample than are persons in the other strata.

This type of sampling can be used to create a more efficient sample design with more statistical power to detect key differences within a population than a simple random sample design or a proportionate stratified sample design. An example of a difference within a population is the comparison of older and younger persons with respect to some characteristic, such as having health insurance. However, a disproportionate allocation can also produce some results that are much more inefficient than a simple random sample or a proportionate stratified sample design.

Disproportionate allocation to strata as a technique can be more efficient than a simple random sample design. Efficiency is determined by whether the sample variances are smaller or larger than they would have been if the same number of cases had been sampled using a simple random sample.

Researchers use disproportionate allocation to strata in order to increase the number of persons with important characteristics within their final study sample and to increase the efficiency of the sample design over simple random sampling. When making estimates using a sample that has used disproportionate allocation to strata sampling, it is important to control for the differences in the probabilities of selection into the sample. Persons from some strata will have been more likely to be included than persons from other strata. To accomplish this task, survey weights are used to adjust each person for their probability of selection into the sample when making estimates of specific characteristics for the entire population.

Disproportionate allocation to strata can make some estimates more (or less) efficient than if the same number of cases had been selected using simple random sampling. Efficiency is gained to the extent that the variables used to stratify the target population are related to the characteristic being studied. For example, when stratifying a health insurance survey by age into two strata—those 65 years of age and older and those under 65 years—the outcome variable of interest, "health insurance coverage," is strongly related to the variable used to stratify. People 65 years of age and over are much more likely to be insured than those under 65 years. The same is true for case-control studies where the condition of interest is used to stratify the target population and the resulting sample is more efficient for studying differences between those with a condition and those that are known to have the condition than would have been possible through a simple random sample of the population.

By the same token, the survey that was more effective for some estimates as a result of stratification may be less efficient for other estimates than a simple random sample would have been. For example, when studying political party preference using a survey stratified by age—65 years and over versus 18- to 64-year-olds—will not yield nearly as efficient an estimate as it did for health insurance coverage, because party preference is not as correlated with being over 64 years as health insurance coverage is. This situation also varies by how much more likely people in one stratum were to be selected into the sample. The worst case scenario is when the strata are completely unrelated to the variables being examined and the two (or more) strata were selected with vastly different probabilities of selection (say 1 in 10 in one strata and 1 in 100,000 in the other). In this case, disproportionate allocation to strata produces an inefficient sample design.

A key aspect of disproportionate allocation to strata is the importance of the estimates for the entire sample versus the estimates for population domains. In our example we have decided to oversample older adults relative to younger adults so that we can compare these two domains with respect to a characteristic such as health insurance. The study will, however, also produce an estimate of health insurance coverage for the entire population, and the oversampling of older adults introduces unequal weights into the sample design that in most cases will reduce the precision of the estimates for the entire population. This type of trade-off needs to be carefully considered at the design stage.

Michael Davern

See also Case-Control Study; Design Effect (*deff*); Optimal Allocation; Proportional Allocation to Strata; Simple Random Sample; Stratified Sampling; Weighting

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury.

Do-Not-Call (DNC) Registries

A do-not-call (DNC) registry is a listing of people who have requested that they not receive any telephone calls from telemarketers on their residential telephone number(s). In the United States, these registries are maintained by the federal government and by government agencies in the majority of states. Survey research organizations are *not* prohibited from calling the numbers on the DNC registries, but many citizens fail to understand this, which causes them to be less willing to agree to participate when sampled for a telephone survey.

The birth of DNC registries has its roots firmly in the groundswell of public reaction in the United States to certain practices of the telemarketing industry that began in the 1980s and escalated considerably until 2003, when the national registry was implemented. The sheer volume of calls, telemarketers' penchant for calling in early evening hours (i.e., dinner time), and their abusive use of predictive dialing equipment all served to overwhelm the repeated attempts of the Direct Marketing Association and the American Telemarketing Association to gain industry agreement and compliance on self-policing measures. This failure of industry standards led to a snowballing of political sentiment that in turn led to legislation in 40 states and ultimately an umbrella National Do Not Call Registry overseen by the Federal Communication Commission (FCC) and the Federal Trade Commission (FTC).

The do-not-call registry guidelines and restrictions generally do not directly affect the survey research industry—in fact, survey research is *specifically exempted* from the law. So, only in rare cases does the legislation ever have the chance of affecting the research industry. However, there is always the potential for interference by an overzealous prosecutor in some potential gray area, and consumers often do not recognize the difference between a telemarketing call and a request to participate in a survey. Indirectly, and through public ignorance, the DNC registries cause some sampled residents whose number is listed on the registry to refuse to participate in telephone surveys, at least in part because they mistakenly believe that survey researchers are restricted from calling their number.

Most citizens associate the DNC with the federal registry that was established in 2003. However, the first state-level DNC was put in place more than 15 years earlier, in Florida in 1987. The federal DNC was established under the 1994 Telemarketing and Consumer Fraud and Abuse Prevention Act directing the FTC, and the 1991 Telephone Consumer Protection Act providing authority to the FCC, to establish such a registry. Proposed rules and public discussions involving the establishment of the DNC began in 1999, with the first registrations taken in October 2003. In total, 40 states enacted their own DNC laws by the time the federal DNC began operation. In most cases, each state set up their own registration procedures for residents of their respective jurisdictions. During the ensuing 5 years since the federal DNC was established, all but one state (Pennsylvania) has folded their registries into the federal database. It is worth noting that the Direct

Marketing Association, the largest telemarketing industry group, had their own do-not-call service, the Telephone Preference Service (TPS), for many years prior to establishment of the federal DNC; the likelihood is that they will follow the lead of the states, ceding their efforts to the federal DNC.

As of the fall of 2007, the U.S. federal DNC totaled approximately 150 million distinct telephone number listings. Only about 55% of these telephone numbers fell within the normal telephone landline random-digit dialing (RDD) sampling frame, another 37% were cellular phone numbers, with the remainder being likely business numbers.

There is one area where the federal DNC can directly affect survey research. In some cases, a telephone survey sampling design may utilize compiled lists of various types to supplement or as an aid in stratifying the sample frame. Most list compilers and vendors are part of the direct marketing world rather than the survey research world. Hence because of their internal business rules, they may eliminate records with telephone numbers that are also on the DNC. Consequently, a telephone survey researcher who buys a sample from such a vendor must question the supplier carefully and in many cases be required to sign a release before the vendor will supply those records to the survey researcher.

There is not much empirical evidence on what effect the DNC registries have had on telephone survey response rates. One study, using large national RDD survey databases compiled in late 2003 and early 2004, indicated that those households whose telephone numbers were listed on the national DNC actually were *more likely to agree to participate in a telephone survey* when they were sampled than were households whose home number was not listed on the registry.

Dale W. Kulp

See also Federal Communications Commission (FCC) Regulations; Federal Trade Commission (FTC) Regulations; Predictive Dialing; Telemarketing; Telephone Consumer Protection Act of 1991; Telephone Surveys

Further Readings

Federal Communications Commission, National Do-Not-Call registry: http://www.fcc.gov/cgb/donotcall

Lavrakas, P. J. (2004, May). *Will a perfect storm of cellular forces sink RDD sampling?* Paper presented at the 59th annual conference of the American Association for Public Opinion Research Conference, Phoenix, AZ.

National Do Not Call Registry: https://www.donotcall.gov

Don't Knows (DKs)

"Don't Know" responses (DKs) occur when respondents report that they do not know or are uncertain about the answer to a survey question, whether it be about their behavior, attitudes, beliefs, perceptions, or a factual question. Don't Knows are often considered to be nonsubstantive responses and a form of item nonresponse. A high level of DKs is sometimes used as an indicator of poor data quality. DKs also create problems for statistical analysis. Researchers often treat DKs as missing data and often exclude cases in which respondents said "Don't know" from their analyses.

Respondents may say they don't know in response to a survey question for a number of reasons. They may genuinely not have an opinion or belief to report. Alternatively, satisficing theory suggests that respondents may report that they don't know because it is a strategy for providing acceptable (satisfactory) answers to survey questions without going through the mental processes necessary to provide a careful response. Finally, some respondents may say they don't know to avoid answering an uncomfortable, embarrassing, or politically charged question.

Don't Know responses are one example of a larger category of no-opinion responses, which reflect that respondents do not have a judgment to report. "Don't know" is sometimes included with the list of substantive response options that are offered to respondents in survey questions that use a closed-ended response format—for example, *or don't you know?* When presented in a survey question with a list of response options that does not include an explicit don't know option, respondents may volunteer "don't know" responses to interviewers in telephone and face-to-face surveys and may write them in on self-administered surveys.

Thus, respondents can report that they don't know the answer to a survey question regardless of whether such an option is explicitly offered to them. However, explicitly including such an option to respondents dramatically affects the proportion of respondents who say they don't know. The inclusion of an explicit

Don't Know response option has been found to substantially increase (from 5 to as much as 30 percentage points) the proportion of respondents who report that they don't know, particularly for questions about issues with which respondents may not be familiar.

Because including an explicit Don't Know option can have a dramatic impact on responses, the decision about whether to explicitly offer such a response option is a very important one for researchers when creating a survey instrument. Two perspectives—nonattitude and satisficing—provide competing theoretical arguments about this decision.

The nonattitude perspective suggests that respondents who genuinely do not know an answer nevertheless may choose a substantive response option when no other option is available. The nonattitude perspective comes from Philip Converse's observation that survey interviews may exert implicit pressure on respondents to appear to have an opinion on a wide range of topics. When respondents are faced with a question to which they genuinely do not know the answer, many may be uncomfortable admitting that they know little about the topic or that they do not know the answer, and this may be particularly true when multiple questions for which they are uninformed are included in a survey interview. Respondents who do not have attitudes on an issue may respond to a question about the issue essentially by randomly selecting responses from among the choices offered. Including an explicit Don't Know response option would provide these respondents with a way to accurately report that they do not know how to answer the question.

In contrast, as noted previously, the satisficing perspective suggests that respondents may choose an explicitly offered Don't Know response option as an alternative to completing the work necessary to choose a substantive response that they would otherwise be able to provide. Thus, the satisficing perspective suggests that Don't Know responses should not always be viewed as accurate reports of nonattitudes. This perspective on Don't Know responding is based on the argument that answering survey questions is a demanding cognitive task. When answering each question in a survey, respondents must understand and interpret the question, search their memory for relevant information, integrate that information into an opinion, and translate that opinion into an understandable response. This "work" may overwhelm respondents' abilities or motivation. In such situations, some respondents may satisfice by seeking out ways to avoid doing this work while still appearing as if they are carrying on a survey interview appropriately. When respondents satisfice, they look for a cue in the question suggesting how to do so. An explicitly offered Don't Know response option provides such a cue, allowing respondents who are otherwise disposed to satisfice to do so by saying, "Don't know." If a Don't Know option was not offered (and the question provided no other cue about how to satisfice), these respondents might be pushed to do the cognitive work necessary to carefully answer the survey question.

Evidence about why respondents choose to report that they do not know generally supports the satisficing perspective. Omitting Don't Know response options from survey questions does not appear to substantially reduce data quality. There is little evidence that explicitly offered Don't Know response options provide an advantage to researchers.

Allyson Holbrook

See also Forced Choice; Missing Data; Nonattitude; Response Alternatives; Satisficing

Further Readings

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly*, *47*, 528–546.

Converse, J. M. (1976). Predicting don't know in the polls. *Public Opinion Quarterly*, *40*, 515–530.

Faulkenberry, G. D., & Mason, R. (1978). Characteristics of don't know and don't know response groups. *Public Opinion Quarterly*, *42*, 533–543.

Gilljam, M., & Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly*, *57*, 348–357.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., et al. (2002). The impact of "Don't know" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*, 371–403.

Double-Barreled Question

A double-barreled question asks about more than one construct in a single survey question. Best practices for questionnaire design discourage use of certain types of questions. Questions with unknown terms or complicated syntax should not be used when designing a questionnaire. Foremost among these recommendations is to avoid double-barreled questions.

The word *and* is a hallmark of a double-barreled question. Double-barreled questions most frequently arise in attitudinal questions. In these types of questions, two attitude targets (e.g., political candidates and policy decisions) are asked as one construct (e.g., *Do you favor candidate X and higher taxes or candidate Y and lower taxes?*). Response formation problems arise when the respondent prefers candidate X and lower taxes or candidate Y and higher taxes. Statements that align two different constructs also are double-barreled (e.g., *Do you agree or disagree with the following statement: Managers in my organization are helpful, but the lack of diversity in the organization is disappointing*). The word *but* plays the role of the conjunction *and*, linking two divergent question constructs into one double-barreled question.

Double-barreled questions require more time for respondents to answer than single-barreled forced choice questions. Comprehension breakdowns are responsible for part of the problems with double-barreled questions. Respondents struggle to understand exactly which construct among the multiple constructs that appear in the question wording is the most important, resulting in higher rates of requests for clarification for double-barreled questions than in single-barreled questions. Breakdowns may also occur when generating a response and in mapping the retrieved or generated response to the response options. As a result, higher rates of item nonresponse and unstable attitudes are likely to occur with double-barreled questions. This also leads to analytic problems and questions of construct validity, as the analyst does not know which "barrel" led to the respondent's answer.

Some double-barreled questions ask about one construct in the question wording, but introduce a second construct through the response options. These questions are sometimes called "one-and-a-half-barreled questions." For example, *Do you agree or disagree with Candidate Z's views on alternative fuels?*, with response options of "Agree," "Agree, and I agree with Candidate Z's stance on tax breaks for hybrid vehicles," "Disagree," introduces the idea of tax benefits from owning hybrid cars only in the response options. As with double-barreled questions, one-and-a-half-barreled questions lead to questions of construct validity. In this example, endorsing "Disagree" can be seen as disagreeing with the candidate's views on alternative fuels, tax benefits for hybrid vehicles, or both.

Turning a double-barreled question into two forced choice questions or two separate statements are common repairs for this questionnaire design problem. Many double-barreled questions mimic forced choice questions but differ in question wording. In a forced choice question, the respondent is asked to choose between two constructs. However, each construct in a forced choice question is asked as a separate idea (e.g., Candidate A versus Candidate B; higher taxes versus lower taxes). Hence, repairs for double-barreled questions can be accomplished by identifying the multiple constructs in the question, deleting irrelevant constructs, and separating relevant constructs into two or more questions.

Kristen Olson

See also Comprehension; Forced Choice; Measurement Error; Questionnaire Design; Questionnaire-Related Error

Further Readings

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*, 390–399.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions.* San Francisco: Jossey-Bass.

Double Negative

A *double negative* refers to the use of two negatives in one statement or question. In questionnaire design, this is almost always a situation to be avoided. A double-negative usually creates an unnecessary amount of confusion in the mind of the respondent and makes it nearly impossible for the researcher to accurately determine what respondents were agreeing or disagreeing to.

Such a question can increase item nonresponse by increasing the percentage of respondents unable to understand the question. A more insidious problem is an increase in the number of responses from people who have misunderstood the question and responded based on that misunderstanding. Both item nonresponse and misunderstanding contribute to the type of survey error know as *measurement error*.

Sometimes the questions sound sufficiently confusing that no serious researcher would include them in a questionnaire. For example, *Are you likely to vote for or against a candidate who is opposed to the proposed ordinance to prohibit smoking in public places?*

However, the second of the two negations often appears in the answer options. It is in these cases when a double negative question may seem perfectly clear to researchers, particularly those who are caught up in issues of the moment. For example, members of the media who support release of the names of jury members might wish to ask:

> *Please tell me whether you agree or disagree with this statement: "The names of jury members in capital trials should not be withheld from the media."*

Perhaps the most infamous example of a double-negative occurred in November 1992 in a survey conducted by the prestigious Roper Organization, a respected survey research center founded in 1947. Roper was commissioned by the American Jewish Committee to conduct a survey of adults in the United States to measure public attitudes and beliefs about Jews. The following question slipped through the usual quality control steps:

> *The term* Holocaust *usually refers to the killing of millions of Jews in Nazi death camps during World War II. Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?*

The published results of the survey implied that one third of adults in the United States felt it was possible that the Holocaust never occurred. The outrage and confusion resulting from the release of the study results prompted several other studies conducted with the specific aim of avoiding the double-negative problem. The other studies worded their survey questions to avoid double negatives. Results from these studies revealed that fewer than 10% of the population felt it was possible the Holocaust had never occurred.

The double-negative problem is difficult to avoid in questionnaire design for attitude and opinion surveys because researchers are often presenting choices to respondents that represent positives and negatives. One approach to avoiding double negations is to make every effort to present the issue without using any negative statement. If the double negative must be used, cognitive interviewing should be employed during pretesting to ensure that respondents have a clear understanding of the question.

James Wolf

See also Cognitive Interviewing; Measurement Error; Missing Data; Questionnaire-Related Error

Further Readings

Ladd, E. C. (1994, July/August). The Holocaust poll error: A modern cautionary tale. *The Public Perspective*, *5*(5), 3.

Drop-Down Menus

Drop-down menus are often used in Web surveys and are one of the basic form elements in HTML (hypertext markup language) used for closed-ended survey questions, in addition to radio buttons and check boxes. They are also referred to as "drop-down lists" and "drop-down boxes."

Drop-down menus commonly display a single option that can be left blank, and other response options become visible after clicking on the side of the box. A response is selected by then clicking on one of the displayed choices, and multiple selections can be allowed (i.e., "check all that apply"). They are suitable and used for long categorical lists, such as lists of states or institutions, but they can be used for shorter and ordinal lists. While radio buttons and drop-down menus can fulfill the same purpose, there are some key differences. Drop-down menus take less space, as all the options do not need to be visible at all times. They also require two clicks, instead of a single click, to select a response option.

Experimental studies show no difference between radio buttons and drop-down menus in terms of time and break-off rates, but find higher rates of item nonresponse and nonsubstantive responses in drop-down menus.

Andy Peytchev

See also Check All That Apply; Closed-Ended Question; Radio Buttons; Web Survey

Further Readings

Couper, M. P., Tourangeau, R., Conrad, F., & Crawford, S. (2004). What they see is what we get: Response options for Web surveys. *Social Science Computer Review*, *22*(1), 111–127.

Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in Web surveys. *Social Science Computer Review*, *20*(4), 471–484.

Dual-Frame Sampling

Dual-frame sampling designs are a subset of multiple-frame designs in which units within the population of interest are selected via independent probability samples taken from each of two frames. These two frames make up the population of interest, and they typically overlap. The dual-frame sampling approach is often useful when the amount of undercoverage from using a single frame is substantially improved by the introduction of two (or more) frames. The degree of overlap in the two frames is usually not known a priori to the sampling, but should this information be available, estimates of the amount of undercoverage to be expected from the dual-frame approach can be assessed more accurately. The resulting estimates from each of the two frames in the dual-frame sampling design are combined to form a single composite dual-frame estimate of the population parameter(s) of interest. A generic figure illustrating the basic structure of a two-frame design is provided in Figure 1.

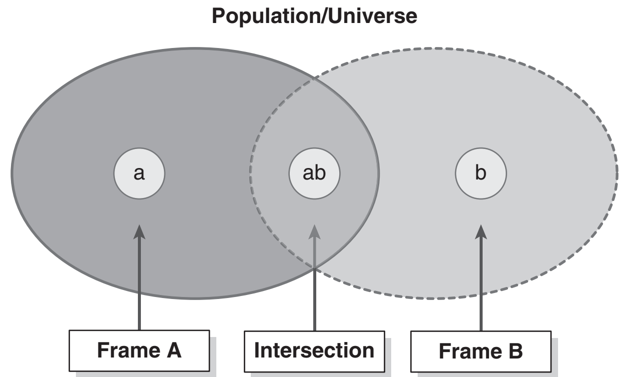

Figure 1 Illustration of two frames for a dual-frame sampling design

Considering this figure, we can see that there are three possible "overlap" situations that may occur when using two frames in the sampling design, including the following:

1. Illustrated in Figure 1 is the circumstance in which neither of the two frames is completely included in the other, implying that Frame A and Frame B have some degree of overlap (i.e., like cell phone and landline phone ownership). This approach serves to improve the overall coverage of the target population, thus reducing undercoverage; this situation is very common for dual-frame designs in practice. Another spin on this approach comes when estimates from a rare population are desired. For example, using random-digit dialing (RDD) to survey the state to estimate the quality of life of breast cancer survivors one year beyond their cancer diagnosis is possible through the use of a health eligibility screener—however, within a given state, the proportion of adult citizens who are one-year breast cancer survivors may be small, making the screener approach alone prohibitively expensive. The State Cancer Registry, however, provides a list of those diagnosed with cancer and is considered "complete" somewhere around 2 years post-diagnosis. So using this frame at the one-year point would certainly be accompanied by a degree of undercoverage and may contain errors in diagnosis, in general but it would include more individuals from the target population of interest. Using a dual-frame approach with an RDD frame with a health screener along with the cancer registry frame may be a more viable and precise approach for estimating the quality of life parameter of interest.

2. Not illustrated in Figure 1 is the circumstance in which Frame A is a complete subset of Frame B (i.e., a rare segment of the population, like homeless, institutionalized, or members of a health maintenance organization who were prescribed a particular type of drug). In this case, Frame B may provide complete coverage of the population frame (i.e., complete household address list for customers within a business district of a large retail corporation), while Frame A may consist of a subset of population units from

Frame B (i.e., an email register of frequent shoppers). If the company wanted to select a random sample of customers, it may be more expensive to sample solely from Frame B based on costs associated with in-person or mailed surveys; to reduce expected costs, a sample from Frame B could be augmented with a sample from Frame A, since emailed versions of the survey would be less expensive to administer than mailed versions.

3. Also not illustrated in Figure 1 is the circumstance in which Frame A and Frame B have no overlap (i.e., list frame of hospital addresses in the northern region and a telephone directory of hospitals in the southern region of the country). In this case, the dual-frame sampling design would simplify to a stratified sampling design in which two strata (northern and southern regions) use different mechanisms for sampling (using addresses versus phone numbers, for example).

A very common estimator of a population total based on a dual-frame sampling design is the *composite* estimator first proposed by H. O. Hartley. This estimator combines estimates of regions (a) and (b) of Figure 1 with a linear combination of two estimates of region 2 derived from the probability samples taken from frames A and B, respectively. Specifically, the estimator is given by

$$\hat{Y} = \hat{Y}_a + \hat{Y}_b + \lambda\hat{Y}_{ab} + (1-\lambda)\hat{Y}_{ba}, \quad (1)$$

where $\hat{Y}_a$ is the estimate of region (a) derived using elements of the first sample that only belong to Frame A, while $\hat{Y}_b$ is the estimate of region (b) derived using the elements from the second sample that belong only in Frame B; similarly, $\hat{Y}_{ab}$ is the estimate of region (ab) derived using the portion of the sample from Frame A that also belongs to Frame B and $\hat{Y}_{ba}$ is the estimate of region (ab) derived from the portion of the sample from Frame B that also belongs to Frame A. The mixing parameter, λ, lies between 0 and 1 and is optimally derived to minimize the variance of the population total estimate as a function of the costs associated with sampling in each of the two frames.

Another form of this estimator was proposed by W. A. Fuller and L. F. Burmeister and makes use of estimates of the population size of the overlap in Frames A and B (i.e., size of region (ab) in Figure 1) but requires estimating two parameters for combining information from the two frame samples. While the form of the estimators is straightforward, the optimal parameters for combining information across the two frames must be estimated separately for each possible outcome variable Y for which population totals are desired. C. J. Skinner and J. N. K. Rao proposed an alternative *pseudo-maximum likelihood* estimator that relies on a single set of global weights that can be universally applied for any outcome variable for which population total estimates are desired. Regardless of the estimator used, there are some overall advantages and disadvantages of dual-frame sampling designs, including the following:

Advantage 1: Increased coverage than what is afforded using only a single frame—using multiple frames generally removes undercoverage bias that is accompanied with only one (incomplete) frame.

Advantage 2: Flexibility in sampling designs per frame (i.e., stratified design in one frame and simple random sample in another, depending on the type of frame).

Advantage 3: Improved coverage of rare populations at lower cost—by using a second frame for the "rare" population units, screeners on the larger, more general frame can be reduced and optimal cost allocation models can be derived based on the distribution of rare population elements.

Disadvantage 1: More complicated estimators and weighting scheme—the composite estimator proposed by Hartley shown in Equation 1, for example, is one such composite dual-frame estimator that requires computation of at least four estimates of the population parameter that will be combined to form the overall dual-frame estimator. Additionally, nonresponse adjustments and other weighting adjustments may be done on the overall composite weights, or separately per frame, then used in the composite estimator. If done separately, control totals and other aspects of the weighting adjustments will be required for each type of frame.

Disadvantage 2: Requires estimation of the "mixing" parameter to deal with how the two estimates of the intersection of the two frames will be combined—estimates usually need to be derived iteratively from the data.

Disadvantage 3: Dual frame designs may have more administrative costs associated with them—that is, two types of surveying required or screeners required for the two frames; different types of collection required to be compatible with frame type (i.e., when using an

Table 1 Example for computing dual-frame estimate for the total purchases made by "rewards program" members based on samples from telephone and email frames

Sample Unit	*Selected from Frame:*	*In the Overlap?*	*Used in:*	*Sampling Weight*	*Annual Purchases for Selected Unit*
1	A	No	$\hat{Y}_a$	16	\$354.39
2	A	No	$\hat{Y}_a$	16	\$205.76
3	A	No	$\hat{Y}_a$	16	\$329.39
			$\hat{Y}_a$ **Estimate: \$14,232.40**		
4	A	Yes	$\hat{Y}_{ab}$	16	\$255.53
5	A	Yes	$\hat{Y}_{ab}$	16	\$264.48
			$\hat{Y}_{ab}$ **Estimate: \$8,320.13**		
1	B	No	$\hat{Y}_b$	10	\$408.70
2	B	No	$\hat{Y}_b$	10	\$415.37
3	B	No	$\hat{Y}_b$	10	\$479.48
4	B	No	$\hat{Y}_b$	10	\$437.05
5	B	No	$\hat{Y}_b$	10	\$311.97
6	B	No	$\hat{Y}_b$	10	\$360.17
			$\hat{Y}_b$ **Estimate: \$24,127.44**		
7	B	Yes	$\hat{Y}_{ba}$	10	\$357.44
8	B	Yes	$\hat{Y}_{ba}$	10	\$394.40
9	B	Yes	$\hat{Y}_{ba}$	10	\$439.34
10	B	Yes	$\hat{Y}_{ba}$	10	\$494.85
			$\hat{Y}_{ba}$ **Estimate: \$16,860.23**		

area frame, personal interviews may be required as compared to telephone interviews used for the landline frame).

For example, consider the population of interest to customers who are registered for a small company's "rewards program." Interest is given in estimating the total annual purchases from the company for registered customers. Customers can register for the rewards program by providing either their landline telephone number or email address; providing both forms of contact is not necessary, and telephone numbers and email addresses are kept separate for marketing purposes. So, in this case, the size of the overlap in Frame A (telephone list) and Frame B (email list) is unknown unless additional steps are taken to match customer information across the two frames. Assume that the telephone frame (A) has 80 numbers from which a simple random sample of 5 numbers is taken, and assume that the email frame has 100 numbers from which a simple random sample of size 10 is selected. For illustration, assume that the costs associated with data collection in each frame for these sample sizes is similar and that there is no survey nonresponse; also assume that $\lambda = 0.5$. For this example, region (a) of Figure 1 refers to customers who provided only a telephone number, while region (b) refers to those customers who provided only an email address, and, finally, region (ab) refers to those customers who provided both a telephone and email address (i.e., appear in both databases). Using estimates from Table 1, we see that the composite estimate based on this dual-frame sample for the total annual purchases for customers in the rewards program is

$$\hat{Y} = 14{,}232.40 + 24{,}127.44 + (0.5 \times 8{,}320.13) + ((1 - 0.5) \times 16{,}860.23) = \$50{,}950.03.$$

Another application of dual-frame designs that is currently being applied in survey practice involves the revision of traditional random-digit dialing designs that attempt to "cover" the telephone population. Because current trends in cell phone only households have created intolerable levels of undercoverage for single landline frame designs both locally and nationally within the United States, the use of a cell phone only frame in conjunction with a landline telephone number frame has been proposed and implemented in practice. Specifically, J. Michael Brick and others reported in 2006 an application of the dual-frame approach for sampling both cell phone and landline telephones to improve overall coverage of the telephone population that incorporated different weights in the landline, cell phone, and combined estimates to adjust for nonresponse bias.

Trent D. Buskirk

See also Cell Phone Sampling; Coverage; Random-Digit Dialing (RDD); Sampling Frame; Stratified Sampling; Undercoverage; Weighting

Further Readings

Brick, J. M., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, *70*(5), 780–793.

Hartley, H. O. (1974). Multiple frame methodology and selected applications. *Sankhyã: The Indian Journal of Statistics*, *36*(Ser. C, Pt. 3), 99–118.

Lohr, S. L., & Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, *95*(449), 271–280.

Duplication

Duplication refers to the prevalence of an element more than one time on a sampling frame, assuming that the element appears only once in the target population but appears more than once in the sampling frame. As straightforward as this problem and its solution may appear to be, its detection and correction can be complicated, time-consuming, and/or costly.

For example, a sampling frame made up of names of members of a professional organization may list the same person more than once if the professional organization has not cleaned its list well, so that all but one of the variants of the same name are purged—as in trying to narrow down the following names to only one listing: "Joan F. Smithers," "Joan Smathers," "J. F. Smithers," "J. Smythers," and so on. Whether or not all the names in this example are the same person is not certain, but it serves to demonstrate the challenges the issue of duplication raises.

Other times, when there is no real list serving as a sampling frame, such as in random-digit dialing (RDD) telephone sampling, the concept of duplication is somewhat more abstract, since the initial sampling unit in such a survey is a household, and many households can be reached by more than one telephone number. Thus, an RDD frame contains a lot of duplication as it relates to the existence of telephone numbers that reach particular households or businesses. In telephone surveying, this is further complicated by the growth of cell phone ownership, which leads to even more telephone numbers that can reach members of the same household.

The major problem that duplication creates is that it leads to unequal probabilities of selection. Probability samples require that elements have a known, but not necessarily an equal, probability of selection. Thus researchers who want to maintain their probability samples must gather information regarding how many "chances" a selected respondent has to be sampled. With a sampling frame that can be cleaned of duplication, it is incumbent upon the researchers to do this as well as possible before the sample is drawn. Then all elements have similar chances of being selected assuming a simple random sample is drawn. But with other sampling frames, in particular with RDD telephone frames, measures must be taken upon reaching a household or business to determine how many other telephone numbers that exist in the frame could also have reached the household or business. This information can then be used to adjust (weight) the database prior to conducting analyses in order to “correct” the issue of duplication and reduce the potential bias it may create.

Paul J. Lavrakas

See also Cell Phone Sampling; Elements; Probability of Selection; Random-Digit Dialing (RDD); Sampling Frame; Target Population; Weighting

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Merkle, D. M., & Langer, G. (2008). How too little can give you a little too much: Determining the number of household phone lines in RDD surveys. *Public Opinion Quarterly*, *72*(1), 114–124.

E

e

e is a term used in the calculation of survey response rates; it represents the proportion of sampled cases with unknown eligibility that are estimated to be eligible cases.

To determine response and other outcome rates for surveys, all cases in the sample first need to be classified into one of four categories: (1) completed cases; (2) eligible cases, no interview (nonrespondents); (3) cases of unknown eligibility, no interview; and (4) not eligible cases (out of sample). Then the eligibility status of the unknown cases needs to be estimated. The proportion of unknown cases that is estimated to be nonrespondents (i.e., eligible cases with no interviews) is known as the *e*-rate and is represented as *e* in equations. For example, in the formula for Response Rate 3, according to the standards of the American Association for Public Opinion Research (AAPOR), the response rate is the number of complete interviews (I) divided by the number of complete interviews (I) plus the number of partial interviews (P), plus the number of nonrespondents due to refusals (R), noncontact (NC), and other reasons (O), plus the number of unknown known cases (unknown if household (UH) and other unknowns (UO)) times their estimated eligibility rate (*e*):

$$RR3 = I/(I+P) + (R+NC+O) + e(UH+UO).$$

So if 55% of the cases in the unknown category were estimated to be eligible cases, *e* would be .55, and 55% of the unknown cases would appear in the base (denominator). The 45% estimated not to be eligible would be excluded from the calculation of this response rate.

In estimating *e*, AAPOR requires that "one must be guided by the best available scientific information" and "one must not select a proportion in order to boost the response rate." AAPOR has documented eight general methods for estimating the eligibility rate:

1. Minimum and maximum allocations
2. The proportional allocation or CASRO method
3. Allocation based on disposition codes
4. Survival methods using either (a) the number of attempts only or (b) the number of attempts and other attributes of cases
5. Calculations of the number/proportion of eligible population compared to same in the realized sample
6. Contacting databases or information sources, such as telephone companies
7. Linking to other records, and
8. Continued interviewing, especially after the close of the data collection period

Cases of unknown eligibility are rare in some types of surveys, such as in-person, area probability samples, but are common in other surveys, such as random-digit dialing (RDD) samples, and mail and Internet samples. In RDD samples in general and especially in RDD samples with minimal callbacks, the number of cases of unknown eligibility (due to "ring–no answer" and "busy" outcomes) will be appreciable, and as a result calculating the response rate will be notably influenced by estimates of *e*. The same will be the case in mail and Internet surveys because of those sampled cases for which nothing is ever heard back by the researcher.

Tom W. Smith

See also American Association for Public Opinion Research (AAPOR); Council of American Survey Research Organizations (CASRO); Eligibility; Ineligible; Response Rates; Standard Definitions; Unknown Eligibility

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Retrieved April 5, 2008, from http://www.aapor.org/responseratesanoverview

Kennedy, C., Keeter, S., & Dimock, M. (2008). A "brute force" estimation of the residency rate for undetermined telephone numbers in an RDD survey. *Public Opinion Quarterly*, *72*(1), 28–39.

Smith, T. W. (2006, January). *Estimating the status of cases with unknown eligibility in telephone surveys*. Paper presented to the Second International Conference on Telephone Survey Methodology, Miami, FL. Retrieved April 11, 2008, from http://www.aapor.org/uploads/Calculating_e.pdf

Ecological Fallacy

The ecological fallacy is a type of faulty reasoning that sometimes is made in the interpretation of results that come from the analysis of aggregate data. This mistake occurs when data that exist at a group or aggregate level are analyzed and interpretations are then made (generalized) as though they automatically apply at the level of the individuals who make up those groups. For example, if a researcher used zip code level census data to determine that the proportion of women in the labor force was inversely correlated with the prevalence of mobile homes in that zip code, it does not necessarily follow that women who live in mobile homes are less likely to be employed than are women who do not live in mobile homes. It is possible that the same relationship between employment and type of housing exists at the level of individuals, but just because it was found to exist at the aggregate level does not assure that it holds at the individual level.

The ecological fallacy can come into play for survey researchers who merge aggregate level data onto their survey data sets—original data that are gathered in a survey from the individual respondents. For example, if a survey of adult residents in a metropolitan area were conducted about race relations, the researchers may want to enhance their analyses by merging zip code or block group census data onto each case in the data set; for example, merging area-level variables such as the percentage of residents in the zip code or block group who are white and the percentage who are black with each survey respondent's individual-level data. These variables can serve many purposes, including being used as statistical controls. They also can allow the researchers to generate new variables by using both the individual-level data gathered in the survey and the aggregate area-level data merged onto the data set; for example, creating a new variable that indicates whether the respondent lives in a zip code or block group in which her or his own race is the majority race. There is nothing inherently wrong with doing any of this, and it does not constitute an instance of committing the ecological fallacy.

Instead, the problem of committing the ecological fallacy occurs when researchers go beyond the precision and applicability of their data to draw conclusions that the data simply do not justify. If the findings are based only on variables that exist at the aggregate level, then no conclusions should be generalized to the individual level. That is not to say the researchers cannot speculate that the same relationships may exist at the individual level. But that is as far as the researchers should go, and it should be labeled explicitly as speculation due to the possibility of the ecological fallacy. If a researcher believes it is important to determine whether the relationship holds at the individual level, it is her or his responsibility to investigate it by conducting a new study that gathers appropriate data at the individual level or by conducting secondary analyses of existing individual-level data.

Paul J. Lavrakas

See also Multi-Level Integrated Database Approach (MIDA)

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

ECONOMIC EXCHANGE THEORY

As it applies to survey research, economic exchange theory provides a possible explanation for why certain types and levels of survey incentives do or do not work to (a) raise the response propensity of a sampled respondent to participate in a survey, (b) improve the quality of the data provided by the respondent, (c) reduce nonresponse bias, and/or (d) lower total survey costs. The central premise in economic exchange theory, as it relates to survey research, is that respondents make at least a partially rational decision about whether or not they will participate in a survey, and the rational part of this decision takes into account the "costs" of participating versus the "benefits" gained from participating. One of those presumed benefits is the value to the respondent of any incentive that may be given her or him by the researchers. Economic exchange theory suggests that if the perceived benefits are equal to or exceed the perceived costs, then the sampled respondent will be positively disposed to participating because she or he will perceive that she or he is being treated equitably by the researchers.

Past research on why people do not want to participate in a survey consistently has shown that "no time" and "no interest" are the two most frequently mentioned reasons for not participating. Consistent anecdotal feedback from survey interviewers indicates that some respondents specifically complain that their time is worth much more than whatever incentive is being offered to them. Thus, some portion of the "No Time/No Interest" constellation of refusal reasons appear to be linked to respondents who rationally are calculating the cost of their time and effort to cooperate versus what is in it for them in return. In most cases the monetary value of survey incentives that have been used throughout the history of survey research has been far too small to approach what most respondents would perceive as an equitable economic exchange for financially offsetting the costs to them for participating.

This reasoning notwithstanding, there are no consistent empirical data that support economic exchange theory as being a driving force in explaining whether certain types of survey incentives will work to raise respondent cooperation and compliance. Part of the reason for this may be that the unconfounded and robust experiments that are required to adequately test economic exchange theory have not been conducted as yet.

That such rigorous research should be conducted is indicated by the results of a provocative study on this topic that was reported by P. M. Biner and H. J. Kidd in the 1990s. This research provided evidence that the manner in which an incentive is framed (i.e., explained to respondents) will affect response rates beyond the mere value of the incentives. Following from personal equity theory, Biner and Kidd used an experimental design that showed that telling sampled respondents that they were being given a small, noncontingent cash incentive as "payment" for participation in the survey—that is, the respondents reasonably were assumed to perceive that they were being "undercompensated" for their time and effort—led to significantly *lower* response rates than telling another group of respondents that the same low-value incentive was being given as a "token of appreciation." In contrast, with larger value noncontingent incentives—especially ones that a respondent likely would view as overcompensation for the time and effort involved in performing the survey task—respondents who were told that they were being given the larger value noncontingent cash incentive as "payment" for participation in the study had significantly *higher* response rates than other randomly assigned respondents who were told that the higher-value incentives were being given as a "token of appreciation."

Many appear to view economic exchange theory as though it is in competition with the long-revered social exchange theory in explaining why incentives work to motivate sampled respondents to cooperate and comply with survey requests. However, it may turn out that the two theoretical perspectives complement each other, but that will await a more rigorous and valid testing of the two theories than heretofore has been conducted.

Paul J. Lavrakas

See also Contingent Incentives; Incentives; Leverage-Saliency Theory; Noncontingent Incentives; Response Propensity; Social Exchange Theory

Further Readings

Biner, P. M., & Kidd, H. J. (1994). The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology and Marketing*, *11*(5), 483–492.
Dillman, D. A. (1978). *Mail and telephone surveys: The total design method for surveys*. New York: Wiley.
Emerson, R. M. (1990). Social exchange theory. In M. Rosenburg & R. H. Turner (Eds.), *Social psychology: Social perspectives* (pp. 30–65). Piscataway, NJ: Transaction.
Goyder, J., Boyer, L., & Martinelli, G. (2006). Integrating exchange and heuristic theories of survey nonresponse. *Bulletin de Méthodologie Sociologique*, *92*, 28–44.
Seta, J. J., & Seta, C. E. (1982). Personal equity: An intrapersonal comparator system analysis of reward value. *Journal of Personality and Social Psychology*, *43*, 222–235.
Wilden, A. (2003). *System and structure: Essays in communications and exchange* (2nd ed.). London: Tavistock.

Effective Sample Size

Complex sample surveys rarely result in a set of independent and identically distributed observations, because of sample design features such as stratification, clustering, and unequal weighting that are necessary for efficient data collection. Such features affect the resulting variance of survey estimates. The effective sample size is one of several useful measures of the effect of the complex sample design on the resulting precision of the estimates.

A general definition of the *effective sample size* is the sample size for a simple random sample selected with replacement that yields the same variance for an estimate as the variance obtained from the sample design used to collect the data. A simple random sample selected with replacement yields a set of independent observations and is the simplest comparison sample design. It is immediately obvious that there is not a single effective sample size for any one study, since the variance for each outcome, analysis domain, and type of estimate (e.g., mean or regression coefficient) will be different. For example, the effective sample size, n_{eff}, of the mean is the sample size such that $S^2/n_{eff} = Var(\bar{y})$, where S^2 is the population variance of the variable in question and $Var(\bar{y})$ is the variance of the estimate under the sample design used to collect the data. Consequently, $n_{eff} = S^2/Var(\bar{y})$.

A related concept is the *design effect (deff)*, which is the ratio of the variance under the sample design used to collect the data to the variance of a simple random sample selected with replacement of the same sample size. Assuming that the sampling fraction for the simple random sample is small, the design effect of the mean is $DEFF = Var(\bar{y})/(S^2/n)$, where n is the sample size from the sample design used to collect the data. Thus, we see that $n_{eff} = n/DEFF$. This latter expression is often used as the definition of the effective sample size. However, the definition presented herein more directly relates to the underlying concept of the effective sample size, whereas its relationship to the $DEFF$ is a consequence of the concept.

To better understand the effective sample size, it is useful to consider the four major aspects of complex sample design that impact the variance of an estimate and hence the $DEFF$ and n_{eff}.

1. *Stratification*. Stratification is the process of dividing the population into mutually exclusive and exhaustive groups and then selecting a separate independent sample from each stratum. When the observations within each stratum are more homogenous than those between the strata, the variance of the resulting estimate will be reduced. If the observations are approximately linearly related to the stratification variable, then the variance of the mean will be reduced by approximately $D_S = (1 - r^2)$, where r is the correlation between the variable under study and the stratification variable.

2. *Clustering*. When clusters, or groups, of observations are selected together rather than single observations, the variance of an estimate is usually increased, since the observations within a cluster are most often positively correlated. In a two-stage sample design, where clusters are sampled first followed by individual observations within each cluster, the amount of increase in the variance of the estimated mean is approximately $D_C = 1 + (m-1)\rho_y$, where m is the number of observations selected per cluster from the analysis domain and ρ_y is the intracluster correlation between two observations in a cluster. This model assumes that the same number of observations is selected within each cluster and that there is a constant intracluster correlation within all clusters. For regression coefficients, the inflation, or possible deflation, in variance is approximately $D_C = 1 + (m-1)\rho_y\rho_x$ where ρ_y and ρ_x are the intracluster correlation coefficients for the dependent

variable and the independent variable, respectively. For certain designs and regression models, it is possible for ρ_x to be negative, resulting in a decrease in the variance of the estimated coefficient.

3. *Unequal weighting*. When the sample is selected with unequal probabilities, the variance of the estimated mean is increased above that of an equal probability sample of the same sample size due to the variability in the weights unless the selection probabilities are approximately proportional to the values of the associated observations or otherwise optimally allocated to minimize the variance. The amount of this increase, often called the "effect of unequal weighting," is approximately $D_W = n\sum_i w_i^2/(\sum_i w_i)^2$, where w_i is the sampling weight for the i^{th} observation. When the weights are all equal, D_W is minimized and equal to 1.

4. *Finite population sampling*. When selecting a sample from a finite population, the variance is reduced when the sample size becomes a substantial fraction of the population size. For a simple random sample selected without replacement, the variance is reduced by the finite population correction factor (fpc) equal to $(1 - f)$, where f, the sampling fraction, is n/N and N is the population size. It is often the case that f is quite small and can be ignored. In the following, we do not consider the fpc since the impact of the other three factors usually dominates the design effect and the effective sample size.

The overall design effect can usually be modeled as $DEFF = D_S D_C D_W$, and the effective sample size by $n_{eff} = n/D_S D_C D_W$, where the impact of sample designs that use stratification, clustering, and unequal weighting are given by D_S, D_C and D_W, respectively. Thus, we observe that effective stratification will increase n_{eff} as r^2 increases. On the other hand, clustering will usually reduce the effective sample size of the mean, since ρ_y is almost always positive. For a regression coefficient, the reduction in n_{eff} due to clustering is often less than that for a mean, since $\rho_y\rho_x \leq \rho_y$ when both intracluster correlations are nonnegative. However, for an independent variable that groups whole clusters together, a so-called between-cluster covariate, $\rho_x = 1$ and ρ_y, again dominates the reduction in n_{eff}. Also, it is possible to make ρ_x negative when the covariate is experimentally manipulated, as in a cross-over design within clusters, and n_{eff} will be increased due to clustering. Finally, the effective sample size for the mean is usually decreased when the variability in the sampling weights is not due to a nearly optimal sample size allocation to minimize the variance or the weights are not inversely proportional to the outcome variable.

In the situation in which clustering has the dominant impact on n_{eff}, there is a very enlightening simple example for the mean estimated from a two-stage design. Consider a two-stage design where 10 $(= m)$ sampling units are selected from each of 50 sampled clusters for a total sample size of 500. If $\rho_y = 1$, then $DEFF = 10$ and $n_{eff} = 50$, the number of clusters. This is the situation where the observations within a cluster are perfectly related and no further information is gained by selecting more than one observation from each cluster. Thus, the effective sample size is the number of clusters. On the other hand, if $\rho_y = 0$, then the observations within each cluster are unrelated, $DEFF = 1$, and $n_{eff} = 500$. This is the situation of independent observations, all of which contribute equal information to the estimate. When ρ_y is between 0 and 1, the effective sample is between 50 and 500.

Last, the effective sample size can be used to estimate power or precision when planning a survey or to calculate the power of an existing survey. In the former situation, n_{eff} can be approximated using the relationships described above for D_S, D_C and D_W combined with data from previous studies to approximate n_{eff} and then used in the appropriate power/precision formula or software package to determine the approximate power or precision. Likewise, the effective sample size can be estimated from an existing survey as $n_{eff} = n/DEFF$ and used in the same way to approximate power or precision.

Rick L. Williams

See also Cluster Sample; Clustering; Design Effects (*deff*); Finite Population Correction (fpc) Factor; Intracluster Homogeneity; ρ (Rho); Stratified Sampling; Weighting

Further Readings

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Neuhaus, J. M., & Segal, M. R. (1993). Design effects for binary regression models fitted to dependent data. *Statistics in Medicine*, *12*, 1259–1268.

Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, *77*, 848–854.

800 Poll

An 800 poll is a one-question unscientific "survey" that is taken by having daily newspaper readers, television viewers, and/or radio listeners call into a toll-free 1-800-number that involves *no cost to the caller*. A different 800-number is given for each "response" that the poll allows the self-selected respondents to choose as their answer to whatever the survey question is. These polls are typically sponsored over a one-day period (or part of a day) by media organizations that produce news. For example, callers who "agree" and those who "disagree" with whichever issue position is being surveyed use separate 800-numbers. It is possible to offer callers more than two answer choices, and thus more than two 800-numbers, but typically the polls utilize only two choices.

Such polls have no scientific standing because there is no way to know what target population is represented by those who choose to dial in. Since this is a nonprobability sample, there is no valid way to calculate the size of the sampling error. Additional threats to their validity include the possibility that the same person will call in more than once. Also, because the response choices normally are limited to two, the question wording and the response choices often are not well crafted.

Nonetheless, they offer a vehicle through which media organizations can provide their audience with a feeling of involvement in the news, since the poll results are typically reported by the news organization within the next day of the poll being conducted. In some cases, the news organizations acknowledge that the poll results they are reporting are unscientific, and in other cases they do not.

With the widespread use of the Internet by the general public, 800 polls have been mostly replaced by similar one-question unscientific surveys on the homepages of news media organizations' Web sites.

It is important to understand that these unscientific 800 polls are entirely unrelated to the scientific use of 800-numbers by some survey organization as a mode of allowing scientifically sampled respondents to "opt into" a mail survey rather than completing a questionnaire and mailing it back. That is, some survey organizations that conduct mail surveys provide their respondents with a toll-free 800-number to call in to complete the questionnaire with a telephone interviewer.

Paul J. Lavrakas

See also Margin of Error; 900 Poll; Nonprobability Sampling; Sampling Error

Election Night Projections

Election night projections are made in the United States on the night of major primaries and major elections to help project the winners of key political races (e.g., president, senator, governor) and issue referenda. A small group of survey research experts make up the Election Night decision teams that make these projected calls. The same basic methods are used to project the winners by the National Election Pool (NEP), a consortium of ABC, CBS, CNN, FOX, NBC, and the Associated Press. Each media organization makes its own projection decisions but relies on common data using this procedure and occasionally supplemented with its own. The method has evolved over the years, but it is essentially the same model developed and implemented in 1967 by Murray Edelman and Warren Mitofsky (then with CBS News) based on their work in probability methods at the U.S. Census Bureau.

Sources of Information

There are four possible sources of information about the election outcome in any given state that are used to make Election Night projections for that state: (1) the actual vote at sample precincts, (2) a statewide exit poll of voters at those precincts, (3) a statewide telephone poll of absentee (early) voters, and (4) the tabulated vote reported by counties throughout the state.

Precinct Sample

Like all estimation, the election night projections start with the sample. In each state where a projection is desired, a sample of voting precincts must be taken. Depending on the size and diversity of a state, a sample of 60 to 120 precincts is taken as the basis of a projection. This procedure for sampling is known as a *stratified systematic sample proportionate to size*.

Each state is broken into three to five geographic areas (geo-strata) based on their historical voting patterns. For example, Figure 1 shows the geographic areas (strata) for New Jersey, while Figure 2 shows an example of New Jersey's historical voting patterns.

A recent past general election in a state is used as the base race for selection for that state. There are three important attributes in choosing the base race: (1) being recent, so as to minimize the geographic boundary changes that may occur from election to

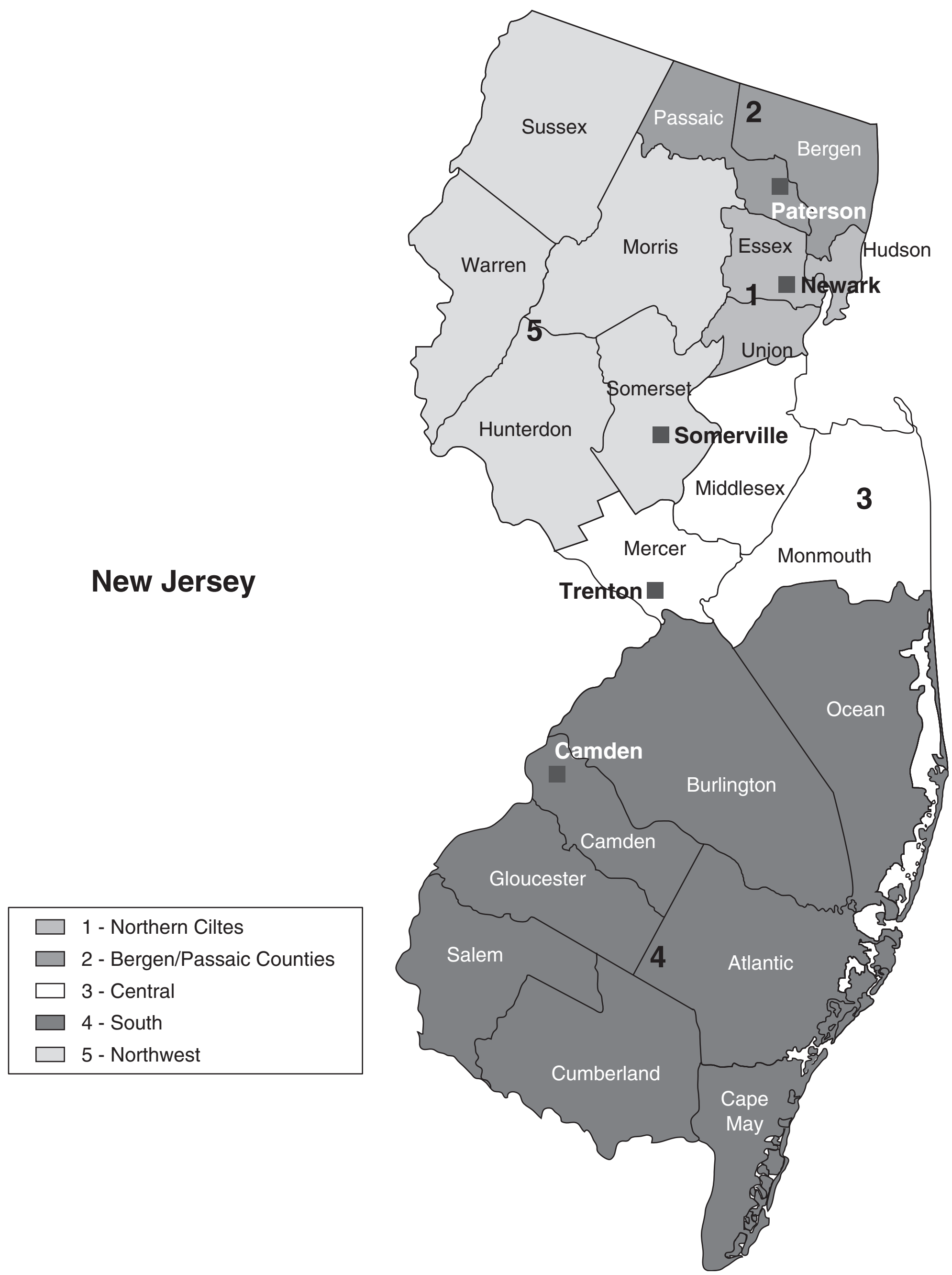

Figure 1 Example of state-level geo-strata (New Jersey)

NEW JERSEY VOTE BY POLITICAL GEOGRAPHY

		Northern Urban	*Bergen/Passaic Counties*	*Central*	*South*	*Northwest*
Proportion of State's 2004 Presidential Vote		*1 (19%)*	*2 (16%)*	*3 (21%)*	*4 (29%)*	*5 (15%)*
2005 GOVERNOR						
CORZINE(D)✓	(53%)	69	56	51	51	40
FORRESTER (R)	(43%)	28	41	44	45	56
2004 PRESIDENT						
KERRY (D)✓	(53%)	66	53	53	51	42
BUSH (R)	(46%)	33	46	47	48	57
2002 SENATE						
LAUTENBERG (D)✓	(54%)	67	55	53	53	39
FORRESTER (R)	(44%)	31	43	44	45	58
2001 GOVERNOR						
MCGREEVEY (D)✓	(56%)	67	56	57	56	42
SCHUNDLER (R)	(42%)	31	43	40	42	55
2000 PRESIDENT						
GORE (D)✓	(56%)	68	56	56	56	42
BUSH (R)	(40%)	29	41	40	41	54
2000 SENATE						
CORZINE (D)✓	(50%)	64	50	51	48	36
FRANKS (R)	(47%)	33	48	46	50	61
50-50 HYPOTHETICAL RACE (President 2004)						
DEMOCRATIC	(50%)	63	50	50	48	39
REPUBLICAN	(50%)	37	50	50	52	61

Figure 2 Example of historical voting patterns (New Jersey)

election; (2) being competitive, so that it shows a reasonable distribution of the vote; and (3) being typical, in that it reflects the ideology of the political parties that the candidates represent.

A listing of all precincts and vote counts in that past election is obtained and geographic codes are added. The precincts are sorted by geographic area and placed in order of total vote. The precincts in each area are then broken down into two separate size groups (usually at the median of accumulated total vote in the geographic strata). The two size groups are used to minimize the bias that can occur when sampling probability proportionate to size. The precincts in each group within a geographic area are ordered by their percentage of Democratic vote in the base race. A sample of precincts is then selected from each group proportionate to the size of their total vote in the base race with an exception in only one state. There are a couple of precincts in New Hampshire that are larger than the interval of selection, and in those cases the precincts are selected with probability of one (1.0). This procedure gives an excellent probability sample of the state for the past race.

It is the job of the research team to properly translate this sample of precincts from a past election to one reflecting the current names and boundaries of the precincts in the upcoming election of interest. This involves talking to county officials to determine if the sampled precinct has had any boundary changes between the selection (base) year and the current year. When there is a boundary change, an effort is made to determine the precinct in the current year that best represents the voters in the precinct that was selected. Once the precincts have been selected, the vote in other past races is obtained to improve the estimation. The current registration in the precinct and its county is also obtained to be used in an estimate of turnout. The selected precincts are then staffed with reporters whose job is to phone in to NEP the actual vote count of the precinct on election night as soon as possible after the polls have closed. They obtain the vote at the actual precinct or at the county where the election results are centrally counted.

The listing of all of the precincts in the state is also used to form a stratification by party. All of the precincts are sorted by their Democratic vote in the *base race*, and different cutoffs are made from High Democratic to Low Democratic.

Exit Poll Data

In most states, an exit poll is also taken. A subsample of between 15 and 60 precincts are selected for the exit poll in each state; the actual number used depends on the importance of the race. The subsampling is done in such a way as to preserve the state's original order and stratification. The interviewers at the exit poll precincts tally the questionnaires three times during the day: morning, afternoon, and about an hour before the polls close in that state. These tallies are the first data used in the projection models.

Early Voting and Absentee Polls

Early and absentee voting is becoming a substantial part of the vote in many states. In Oregon, elections are totally absentee, and in Washington State more than 70% of voters cast absentee ballots. Other states with a large absentee vote are California, Texas, Tennessee, and Arizona. An estimate of the size of the absentee vote is made by an NEP analyst who looks at the size of the absentee vote in past elections along with changes in rules and current requests for absentee ballots in the state. In states with a large absentee vote, it is essential to supplement the exit poll with an absentee poll. These are regular random-digit dialed (RDD) telephone polls with random selection of respondents and multiple callbacks to improve response rates that are conducted before the day of the primary or the election. Respondents are asked if they have already voted or plan to vote before Election Day. If so, they are interviewed using the basic exit poll questionnaire, slightly modified for telephone use. These data are weighted and used to estimate the absentee vote in the state. The sample sizes of these "early voter" RDD surveys range from 400 to 800 depending on the importance of the absentee vote.

In states where there is not an absentee poll, an estimate of the current absentee vote is made when possible using an estimate from past years of the difference between the absentee vote and the Election Day vote. This is, of course, the default method used when the absentee is not sizable or the race in the state is not seen as having enough editorial (news) importance for the substantial expense of an absentee poll.

Vote Tabulation

Once the polls in a state close, a preliminary tabulation is conducted by the state. Many states now put the results by county on their Web sites in a timely manner. But in all states, the vote results, broken down by county, are provided by the Associated Press and are used directly in the election projection models.

Models Used

Models With Precinct Data

There are two basic models used in estimation. The "simple" estimate uses the current proportion of the vote for each candidate in the precinct, averaged over all precincts that have reported in each stratum. At that point, these average proportions are weighted by the size of the strata.

The "ratio" estimate is the average just described using the current vote divided by a similar average of proportions, based only on the past vote in the same precincts that have reported, multiplied by the actual past proportion for the same candidate in that stratum. This stratum estimate is then weighted by the size of the stratum.

There are two stratifications for each of these methods of estimation and hence four estimates: (1) simple geo, (2) simple party, (3) ratio geo, and (4) ratio party. Prior to poll closing all are based on the relative sizes of the strata in a past election. In the two geo estimates after poll closing, the current sizes of the strata are estimated, using an average of the total current vote divided by registration in each reported precinct and multiplied by the total registration in that stratum. In the two party estimates, no current estimate of stratum size is available, so this estimate is used more cautiously when there appear to be any deviations from usual voting trends.

These models are used with the exit poll tallies and then later with the reported votes from the sample precincts. The variance of these estimates uses the standard form of the computation of variance of a stratified random sample.

Models With County Data

Unlike the precinct data, which is a sample of precincts, the county data is an "evolving census" of the vote count. Most counties start with a trickle of precincts approximately a half-hour after the polls close in a state and eventually reach 100% of the precincts, often many hours later. In some states, the absentee vote is counted with the precinct Election Day vote; in other states, it is counted separately and added to the vote, sometimes at the beginning of the night and other times at other irregular intervals. In a few states, at least some of the absentee vote is not counted until days later. An analyst trying to make a projection from the county reports has to be cautious when estimating how much vote has come in at a given time, since the vote count only roughly follows the proportion of precincts reported. In addition, even when a county has 100% of the vote reported, it can still have errors of as much as 0.5%.

A county estimate is made by inflating the votes in the county based on the inverse of the percentage of the precincts reported. Then the counties are cumulated by stratum and inflated to the stratum size to account for counties that have not reported yet. The stratum estimates are then added and the votes percentaged at the state level. An error term on this estimate is formed by using a regression equation that is based on historical data over different time intervals and elections relating the percentage of precincts reporting in a county at a given time to the deviation of the candidate percentages from the final outcome. The estimates of the individual county errors are combined by stratum and then adjusted to the relative sizes of the strata to form an estimate of the error of the overall state estimate.

When the Information Is Used

Prior to Poll Closing

On Election Day, prior to poll closing, the only sources of voting information are the exit and absentee polls. As mentioned, the exit poll interviewers report in three waves, typically in the morning, afternoon, and about an hour before poll closing. At each of these points, both simple and ratio model estimates are made. These are combined with the absentee estimate when available based on the estimated relative size of both types of vote. The estimates are rank ordered based on the variance of each, with the estimate with the smallest variance considered the "best" statewide estimate. The variance of this estimate is computed and a critical value is formed. When the critical value is high enough that the there are only 5 chances in 1,000 ($p = .005$ or less) of being wrong, a call can be made. Judgment always weighs in here because of the possibility of nonrandom error, such as that produced by nonresponse.

Usually there is a pre-election or "prior" estimate of the outcome based on the pre-election polls as combined by political experts. The analyst would always have that prior expectation in the back of her or his mind when considering a projection. Although a combination of the prior and exit poll estimates is not used for projection, the two are composited, using the inverses of their errors, and used for reference.

Shortly After Poll Closing

Once the polls have closed in the state, the vote in the sample of precincts gradually becomes available and is used in the model with the poll data. As each exit poll precinct reports, the actual vote replaces the exit poll tally information in the model. The same stratified estimates are calculated now with the combination of exit poll tallies and actual votes in the precincts and the same decision model employed. When the precinct vote is available in at least eight of the exit poll precincts, it is also possible to estimate the possible error (bias) in the exit poll. This is used as a guide by the Election Night decision teams in evaluating the accuracy of the estimate.

After about an hour, there typically is enough actual vote reported by county to make an estimate from it. This estimate is combined with the best precinct estimate using the inverse of their error so that the more accurate estimate dominates. For the geographic stratification, another estimate, called the "integrated model" becomes possible when there are enough counties reporting in each stratum. This is formed by creating a composite estimate in each stratum of the precinct estimate for the stratum combined with its county estimate. The strata estimates are weighted to their size and summed to obtain an estimate for the state. This estimate is usually the one watched most often by the decision team once there are enough actual votes reported to get an estimate for each stratum.

As the Vote Straggles In

With more than 90% of the vote reported, there is yet another way to view the reported vote. One can look at the outstanding vote by county. For example, if the vote yet to be reported is from counties where the Republican is ahead in the tabulated vote, one would be more confident that the Republican would win. This method, however, has to be used judiciously. Estimates of the outstanding vote in a county depend on assuming that the precincts yet to report are of comparable size to the ones that have already reported and that the candidate vote to be reported is similar to the vote already reported.

Eventually the vote tabulation will reach 100%, but, as already mentioned, even then the state count can be off by as much as a half of one percentage point. Some errors in the vote may persist even after the vote is officially certified by the state months later.

Murray Edelman and Clyde Tucker

See also Exit Polls; National Election Pool (NEP); Probability Proportional to Size (PPS) Sampling; Random-Digit Dialing (RDD); Stratified Sampling; Systematic Sampling; World Association for Public Opinion Research (WAPOR)

Further Readings

Edelman, M., & Merkle, D. M. (1995, May).*The impact of interviewer characteristics and Election Day factors on exit poll data quality*. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Fort Lauderdale, FL.

Merkle, D. M., & Edelman, M. (1997, May). *A comparison of early voters and Election Day voters*. Paper presented at the 52nd Annual Conference of the American Association for Public Opinion Research, Norfolk, VA.

Merkle, D. M., & Edelman, M. (2000). A review of the 1996 exit polls from a total survey error perspective. In P. J. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 68–92). New York: Chatham House/Seven Bridges.

Merkle, D. M., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis." In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 243–258). New York: Wiley.

Mitofsky, W. J., & Edelman, M. (1995). A review of the 1992 VRS exit polls. In P. J. Lavrakas, M. W. Traugott, & P. V. Miller (Eds.), *Presidential polls and the news media* (pp. 81–100). Boulder, CO: Westview.

Mitofsky, W. J., & Edelman, M. (2002). Election night estimation. *Journal of Official Statistics*, *18*(2), 165–179.

ELECTION POLLS

Election polls are surveys that are taken before, during, and after election season and are used to predict and explain election outcomes. The media conduct election polls to satisfy their viewers' and readers' desire for "horse race" journalism and to help editors and reporters plan their coverage of elections and politicians. Candidates and political parties use them for strategic purposes, including fund-raising and helping to position their campaigns in the best possible light. Political scientists and other academics conduct election polls to understand the influence of campaign dynamics on voting behavior.

Election polls employ various survey methods and come in a variety of types. In the United States over the past few decades, most election polls have been random sample telephone polls, drawn from various target populations, such as the entire adult population, registered voters, or so-called likely voters. Few are conducted face to face, but a growing number are conducted via the Internet. Respondents are usually asked a series of standardized questions designed to measure their opinions on issues and candidates. Election polls influence political campaigns in a number of ways and are an important component of the election news coverage. There are several different types of election polls that are conducted at various times during a political campaign, from before the primaries to the day of the election.

Pre-Primary and Primary Surveys

Surveys are conducted early in a campaign to help benchmark baseline information about voter demographics and the public's perceptions of the candidate's image, message, and issue positions. The most useful benchmark questions for a candidate concern name recognition, strengths compared to challengers, and performance while in office (if the candidate is an incumbent). The results of these surveys are circulated within a candidate's campaign organizations and help shape strategy.

These surveys are conducted before and during the season of the primary elections, when campaigns are striving to demonstrate the viability of their candidate. The results are used by the candidates to stimulate fund-raising efforts, and may be leaked to the news media if favorable to the candidate and/or unfavorable to the opponent(s). The value of these pre-primary polls depends on their timing. If conducted too early, respondents may not know enough about a candidate. If conducted too late, the results may have little value to the candidates.

Trial Heat Pre-Election Surveys

The primary focus of these pre-election surveys is to gather trial heat data that essentially take the form, *If the election were held today would you vote for A or B?* The reliability of these questions also depends on the timing of the survey. If they are asked too early in the campaign, these questions are more likely to measure name recognition, not voter intentions. The results are prone to considerable fluctuation related to the changing campaign conditions. For example, a Fox News/Opinion Dynamics trial heat survey conducted in August 2004 gave Democratic presidential candidate John Kerry a 6-point lead over Republican George Bush, but the same survey conducted 2 months later in October gave Bush a 4-point lead, a swing of 10 percentage points. Such surveys are at the heart of what has become known as "horse race" journalism, which refers to the perceived obsession of the news media to focus overly on who is likely to win an election.

Tracking Polls

Tracking polls produce up-to-date estimates of campaign leaders and are typically conducted over the last few weeks of the campaign. They are used by the media to complement its horse race coverage and by candidates to monitor late shifts in support, especially any shifts that may occur after a campaign-staged event or other newsworthy events that may arise. They produce a rolling average estimate derived from daily samples, usually 100–200 interviews each, that typically are aggregated across 3-day periods (e.g., Monday–Tuesday–Wednesday, Tuesday–Wednesday–Thursday, Wednesday–Thursday–Friday). Tracking polls have been criticized for employing inconsistent sampling procedures; they are often conducted only in the evenings, rarely attempt to deal with hard-to-reach respondents, and select respondents based on whoever answers the phone rather than randomly within the household. Tracking polls can be very expensive compared to other pre-election surveys.

Exit Polls

Probably the most controversial of election polls are exit polls. Exit polls have two primary purposes: helping project winners on the evening of Election Day and helping explain election outcomes in the days following the election. These polls consist of interviews of voters as they leave sampled polling places; they are asked a short list of questions concerning vote decision, issue positions, and voter demographics. Exit polls use multiple-stage sampling methods. First, the polling organization randomly samples counties in the states of interest, and then precincts within counties, and then interviewers in the sampled precincts select respondents based on a pre-determined systematic sampling selection method. Respondents usually complete a survey form and place it in a "ballot box" themselves. ABC, CBS, NBC, CNN, Fox, and the Associated Press (AP) conduct the most widely know U.S. exit poll: previously through a consortium called the Voter News Service, and recently through a partnership between Edison Media Research and Mitofsky International. Through the 2006 election, the *Los Angeles Times* offered the main alternative source of national exit poll data but decided not to continue this for the 2008 election. Interviewer bias, differential response, faulty data, and the timing of results reporting are often cited as the major problems associated with exit polls.

The largest benefit of exit polling is that it provides critical data for analyzing the meaning (or so-called mandate) of the election—data that sometimes contradicts the political explanations for a candidate's victory. For example, Republicans often attribute George Bush's 2004 victory over John Kerry as due to the votes of Christian Evangelicals, when in fact exit poll data indicated that middle-class, suburban whites were Bush's largest group supporters.

Influence of Election Polls on Voters and Journalists

The influence of election polls on voters is mixed. On the one hand, some studies indicate that election polls deter some registered voters from voting and sometimes may mislead candidates on the effectiveness of their campaign. Other studies suggest that bandwagon and underdog effects driven by news about pre-election poll standings can change voting preferences within the electorate. Election polls may also bias media coverage toward frontrunners—especially at the time of the primaries, when candidate "viability" often appears to help determine news coverage assignments—leading journalists and thus voters to pay less attention to trailing candidates and more attention to leaders. Election polls have also been argued to influence campaign fund-raising, with donors more likely to give money to frontrunners. On the other hand, studies have consistently found that election polls help inform the public about election and candidates and also get people interested in the political process. Despite the various criticisms, polls can be expected to grow in use and popularity in future elections.

Jamie Patrick Chandler

See also Bandwagon and Underdog Effects; Exit Polls; Horse Race Journalism; Likely Voter; Pollster; Pre-Election Polls; Pre-Primary Polls; Rolling Averages; Trial Heat Question; Tracking Polls

Further Readings

Asher, H. (2007). *Polling and the public: What every citizen should know* (7th ed.). Washington, DC: CQ Press.

Erikson, R, S., & Tedin, K. L. (2007). *American public opinion: Its origins, content, and impact*. New York: Pearson Longman.

Glynn, C. J., Herbst, S., O'Keefe, G. J., Shapiro, R. Y., & Lindeman, M. (2004). *Public opinion* (2nd ed.). Boulder, CO: Westview.

Lavrakas, P. J. (2007, January). *Why our democracy needs accurate national exit polls*. Retrieved April 5, 2008, from http://www.publicopinionpros.com/features/2007/jan/lavrakas.asp

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Elements

Within the context of survey research, an *element* is the basic unit that represents whatever is being sampled and from which survey data are to be gathered. Thus, the elements used in different surveys will depend on the purpose of the survey and may be adults, children, households, employees, businesses, students, teachers, schools, school districts, uniformed personnel, civilian personnel, police districts, libraries, books within libraries, pages within books, or many other things.

Within a target population, all the members of that population are its elements. Within a sampling frame, all the elements from the target population that can be listed constitute the frame. All the elements that are selected for study from the sampling frame make up what is commonly called "the survey sample." However, all the selected elements from which data are gathered also are commonly referred to as the "sample."

Paul J. Lavrakas

See also Sample; Sampling Frame; Sampling Pool; Target Population

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Eligibility

Eligibility refers to whether or not a sampled unit is eligible to have data gathered from it—that is, is the unit part of the survey's target population or is it not? For example, the target population for a survey might be all adults who are 18–34 years of age. As such, if a household is sampled and screened via random-digit dialing and no one living there fits the age criteria, then the household is ineligible. If there is at least one person ages 18–34 years, then the household is eligible. Ultimately, eligibility versus ineligibility is central to the issue of how well a sampling frame and a sample drawn from that frame "cover" the target population and whether or not coverage error results. Eligibility also is linked to survey costs, since samples drawn from frames that contain a large portion of ineligible units are much more costly to process. As straightforward as it may appear to determine eligibility for a survey, it often is not at all easy to do and many mistakes (errors) may occur in the process. Mistakes in determining eligibility may lead to coverage bias.

For example, most surveys have geopolitical boundaries for their samples, as they are not national in scope. In each of these surveys, the target population typically is limited to those residents living within the geopolitical boundaries (e.g., a particular county). If the mode of sampling and data collection is the telephone, as it often is, then some form of geographic screening must be instituted for interviewers to determine the eligibility of the household or person being contacted. In the case of boundaries that are commonly known and well understood by the public (e.g., one's county of residence), eligibility is readily determined without many errors, as long as the respondent does not know what answer will make her or him eligible or ineligible for the survey. (If the respondent knows this in advance of answering the screening questions, some respondents will self-select themselves in or out of the interview erroneously.) On the other hand, if the geographic boundaries that define eligible residency are not well know, (e.g., a school district or a police district), then screening a sample for eligibility via the telephone can be fraught with error. Some people will mistakenly say, "Yes," they live within the boundaries when asked the screening question(s) when in fact they should have said, "No" (errors of commission), and others will say, "No," when they should have said, "Yes" (errors of omission). This will occur even if the screening sequence carefully defines the eligibility boundaries, because many people are "geographically challenged" and will not understand the boundaries they are being asked about.

There are many other criteria than geography that are used to define eligibility in various surveys. For example, a major health survey conducted annually since 1994 for the U.S. government interviews only parents of children ages 19–35 months to learn about the immunization history of the children. (The exact eligibility definition for a household in this survey essentially changes every day of interviewing, since children age each day.) Researchers therefore must pay very careful attention to how eligibility is defined and how it is explained to respondents, whether that be in an interviewer-administered survey (e.g., in person or telephone) or via a self-administered survey (e.g., via mail or Internet).

Eligibility directly affects two important aspects of survey quality. First, coverage problems and possible nonnegligible coverage error will result if eligibility status is not accurately determined throughout a survey. Second, the response rates that are calculated for a survey will be affected by how well the eligibility of the sampled units is determined. The response rates will be depressed if too many ineligible units are deemed erroneously to be eligible, because these units will then be included in the denominator of the fractions used to calculate the response rates when in fact they should be excluded. This too is a complex issue, especially with telephone, mail, and Internet surveys, since the data collection period often ends with the eligibility status of many cases remaining unknown (e.g., those numbers that are always busy during a telephone survey). To accurately calculate response rates, the researchers must make informed judgments about the proportion of these cases with unknown eligibility that are likely to be eligible. This matter is so important that the standard definitions for calculating response rates promoted by the American Association for Public Opinion Research include specific instructions about how to estimate eligibility among those cases that end with a status of "unknown eligibility" (which is referred to as *e*).

Paul J. Lavrakas

See also American Association for Public Opinion Research (AAPOR); Coverage; Coverage Error; *e*; Errors of Commission; Errors of Omission; Geographic Screening; Ineligible; Response Rates; Sampling Frame; Screening; Standard Definitions; Target Population; Unit; Unknown Eligibility

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Kennedy, C., Keeter, S., & Dimock, M. (2008). A "brute force" estimation of the residency rate for undetermined telephone numbers in an RDD survey. *Public Opinion Quarterly*, *72*(1), 28–39.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection and supervision* (2nd ed.). Newbury Park, CA: Sage.

Smith, T. W. (2003). *A review of methods to estimate the status of cases with unknown eligibility*. Retrieved April 5, 2008, from http://www.aapor.org/uploads/Calculating_e.pdf

Email Survey

An email survey is one that sends the survey instrument (e.g., questionnaire) to a respondent via email and most often samples respondents via email. These electronic mail surveys first came into use in the late 1980s, and many scholars at the time thought that they represented the future of survey research. Since then, Web (Internet) surveys have become the predominant model for electronic surveying, because of the relatively poor performance of email surveys in terms of ease of use and response rates.

Email Survey Procedures

Similar to a Web survey, a survey conducted via email most typically uses electronic mail to contact members of the sample. With Web surveys, the user is directed in the contact email to a Web site containing the questionnaire. With email surveys, the contact email contains the survey questionnaire and no survey Website is referenced. Generally, the email survey approach takes one of three forms: (1) a software file attached to the email, (2) an electronic document attached to the email, or (3) questionnaire text embedded in the email itself.

Some commercial vendors have offered survey software that will attach an executable file to each email sent to sample members. The file is downloaded by the user and executed on his or her personal computer; a software program then prompts the user to fill out the questionnaire and records their responses. A more common and simpler approach is to attach a copy of the questionnaire to the email as an electronic document, using a common format such as Microsoft Word. Users open the document, type their responses in the appropriate places, save the document on their computer, and then reattach the document to an email that is sent back to the surveyor. Alternatively, a text or HTML (hypertext markup language) copy of the questionnaire can be embedded directly in the contact email. By hitting "reply" in their email software, users create a copy of the questionnaire into which their responses can be typed. Responses are given either by inserting Xs into a set of brackets by the appropriate response for text emails or by marking radio buttons for HTML emails, and the email is sent to the surveyor.

Emails from respondents are then collected and entered into a database, either by hand or through the use of survey software. With both the electronic document approach and the embedded text approach, users are often given the option of printing off the questionnaire and mailing it back to the surveyor via regular mail. Research indicates that few respondents use this option when it is offered.

Advantages and Disadvantages

Email surveys share many of the same advantages and disadvantages of Web surveys. For example, email surveys are less costly than other survey modes, because of the lack of printing, postage, or interviewer costs, and they enable the surveyor to collect data quite rapidly. Thus, the potential for interviewer effects are eliminated due to self-administration of the survey. Email surveys also share the same coverage issues as Web surveys, in that not everyone has an email address (nor are there national directories of email addresses), as well as the same measurement error issues, arising from the format of the questionnaire appear differently on different computer and software configurations. There is also evidence that respondents to email surveys may differ from respondents to mail surveys in terms of having higher socioeconomic status; similar results have been found for Web surveys.

The main issue with email surveys, especially when compared to the alternative of Web surveys, is the difficulty of use for both the respondent and the surveyor. With the attached software approach, users must not only know how to run an executable file but also be comfortable with running such files on their personal computers. Given that computer viruses are often spread via executable files, it is likely that many respondents are uncomfortable with this approach. Different computer configurations among the target population, as well as variance in file size limitations among email servers, may also hamper the use of this approach.

With the use of attached electronic documents, users must know to save the document on their computer after entering their responses, generate an email to the surveyor, and reattach the saved document to the email before sending. As with attached software, respondents may be wary of opening an electronic document from an unknown sender. While attached documents would appear to offer the advantage of more formatting and graphics compared to embedded text surveys, methodological research studies indicate that email surveys with embedded text yield higher response rates than attaching an electronic document.

With the embedded text approach, users must hit the "reply" button to generate a copy of the survey; not all email programs are set to create a copy of incoming messages (and thus the questionnaire) when users hit the reply button. While email programs that read HTML-based emails are becoming more popular, many individuals still use email programs that can only read text-based emails. Many of these issues could theoretically be overcome with detailed instructions for the respondent, but the increased complexity of the email survey process is a likely cause of the low survey response rates reported by researchers.

An often overlooked issue with email surveys is their confidentiality. Unlike Web use, many businesses and Internet service providers routinely monitor and/or electronically back up their email traffic. Even when businesses closely monitor Web use, such monitoring usually consists of a record of Web sites accessed rather than a record of every keystroke used or Web form submitted. Confidentiality concerns can thus be higher for email surveys, as copies of the respondent's answers may be made on both their personal computer as well as their email server, unlike in Web or mail surveys. In addition, respondents may be wary of replying to an email, fearing that the surveyor is a spammer and is using the survey to verify email addresses for future spamming. This is less of a concern in Web surveys, which generally involve clicking on a hyperlink rather than giving an email address.

Email surveys can also increase the burden for the surveyor. With the electronic document and embedded text approaches, users can drastically alter the survey document by deleting the special symbols (such as brackets) that are used to designate the beginning and end of each item response. When this occurs, the response must be processed by hand because the survey software cannot process the response. Such processing can be resource intensive. For example, M. P. Couper, J. Blair, and T. Triplett found that nearly half of their email survey responses required some sort of clerical action before they could be added to the survey response database.

In order to keep track of respondents, a unique identifier must be included in each document or email; these can easily be deleted by the respondent. Duplicate submissions may then become a problem. Some researchers have tried using email addresses as the unique identifier, but this does not always work in practice. For example, respondents have email forwarded to other accounts, from which they then reply to the survey, making respondent tracking difficult. Respondents may also forward email surveys to individuals outside the target population.

The literature reveals numerous studies comparing email surveys to mail surveys; the majority of studies show higher response rates for mail surveys. Research is mixed as to whether data quality is higher for mail or email surveys. Interestingly, there is little research comparing email surveys to Web surveys.

The Future of Email Surveys

As HTML-based email programs become more popular, email surveys may become more prevalent. Given that their major drawback lies in the complexity of the process, an HTML-based questionnaire could be constructed in an email such that the respondent could fill it out and submit his or her responses without having to access an external Web site. This involves fewer steps than the typical Web survey, for which respondents must click on a URL (uniform resource locater, or Web address) in the contact email to open the survey in a Web browser. Yet until there is more commonality across email programs, which in turn will allow consistency in design, email surveys will likely fill a small niche in the survey researcher's toolbox. Simply put, email surveys face significant issues in terms of how respondents complete and return the instrument, and how the surveyor processes and analyzes completed surveys. Conversely, Web surveys offer the same advantages as email surveys in terms of cost and speed and relatively few of the disadvantages.

Email surveys are likely to be successful when use of a Web survey is impractical and costs preclude the use of other modes. For example, some organizations may limit their workers' access to the Web. Use of an email survey to query employees would be a viable alternative in this situation, especially considering that employees for an organization are likely to use the same email program.

Stephen R. Porter

See also Computer-Assisted Self-Interviewing (CASI); Computerized Self-Administered Questionnaires (CSAQ); HTML Boxes; Internet Surveys; Radio Buttons; Respondent Burden; Web Survey

Further Readings

Best, S. J., & Krueger, B. (2002). New approaches to assessing opinion: The prospects for electronic mail surveys. *International Journal of Public Opinion Research*, *14*(1), 73–92.

Couper, M. P., Blair, J., & Triplett, T. (1999). A comparison of mail and e-mail for a survey of employees in U.S. statistical agencies. *Journal of Official Statistics*, *15*(1), 39–56.

Dommeyer, C. J., & Moriarty, E. (2000). Comparing two forms of an e-mail survey: Embedded vs. attached. *International Journal of Market Research*, *42*(1), 39–50.

Palomares, N. A., & Flanagin, A. J. (2005). The potential of electronic communication and information technologies as research tools: Promise and perils for the future of communication research. In P. J. Kalbfleisch (Ed.), *Communication yearbook* (Vol. 29, pp. 147–185). Mahwah, NJ: Lawrence Erlbaum.

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly*, *62*(3), 378–397.

Sheehan, K. B., & Hoy, M. G. (1999). Using e-mail to survey Internet users in the United States: Methodology and assessment. *Journal of Computer Mediated Communication*, *4*(3), 1–25.

Sheehan, K. B., & McMillan, S. J. (1999). Response variation in email surveys: An exploration. *Journal of Advertising Research*, *39*(4), 45–54.

Encoding

Encoding information is the cognitive process through which experiences are translated into memory. However, for the social sciences, *encoding* often means the process of translating thoughts, ideas, or questions into words. Different phrases and words, definitional and connotative frameworks may conflict given different audiences and contexts. In survey research, the encoding of widely understood and definitive meaning into a question is essential to valid measurement. Researchers must be cognizant of how different groups will interpret (or decode) their questions. A strong survey instrument ensures that the researcher and the respondent share the same understanding of both the questions asked and the answers given. Compounding problems emerge when a respondent is conditioned by the survey questionnaire or must choose between response options with similar meanings.

One example of poor encoding might be translating an idea into a question that the respondents interpret inconsistently as a group, interpret differently from the researcher, or both. For example, a survey question might ask how respondents feel about democracy. In order to interpret responses to this question, a researcher must assume that everyone in the sample shares the same definition of *democracy* as everyone in the sample population and also shares the surveyor's definition. Further, the researcher must receive the respondent's answers with the same understanding that the respondent delivers it. In survey projects, consistent encoding and decoding is essential to relaying the respondents' true responses. Failing to anticipate how different groups of people will interpret the survey instrument will affect both the internal and external validity of a research project.

There are several sources of differential encoding and decoding in survey research. A survey or researcher may ask questions in a second or translated language and may confuse meanings in the second language. Values, ideas, and definitions may differ across cultures or within cultures. Common usage and events may affect how groups understand and the questions or answers in the survey. Also, some ideas are inherently more complex than others. Complex concepts in either the question or the answer can result in multiple interpretations of their meanings.

Encoding may also occur in the questionnaire itself. As sampled people respond to different elements of the questionnaire, they may begin to develop a set pattern based on either the questions asked or the list of possible answers. In this instance, the respondent has been conditioned to the questionnaire itself. She may have encoded a set of answers per question type, established an opinion or mood toward the survey instrument, or may skim over similar questions, believing (correctly or incorrectly) that she has an answer. Hence, the respondent may assume a set of values within the research project that may not accurately reflect her true opinion, preference, actions, and so on that exist outside the particular research study.

Researchers work to minimize encoding and decoding differentials in many different ways. One way to guard against individuals sharing different conceptions is by thoroughly explaining difficult and important parts of the questionnaire. Explicitly stating and defining questions and answers within the study helps homogenize how the respondents respond. Also, the researcher can construct straightforward and simple questions and answers. From these simple responses, the researcher then may develop insight into theoretically deeper questions. Last, to gain external validity, the descriptions in the study should be congruent with other scholarly works, environments, or surveys.

In order to minimize encoding and false reflexive responses based on the questionnaire, the researcher can develop a complex survey instrument that varies questions and responses. When the language, structure, and the order of the questions vary, the subject is less likely to create a fixed response. Such a questionnaire forces the subject to thoughtfully engage the questions and be more apt to respond accurately. Encoding is not necessarily detrimental to research. It is a characteristic of thought and social relations of which researchers must be aware.

Ryan Gibb

See also External Validity; Internal Validity; Questionnaire-Related Error; Reactivity; Respondent-Related Error

Further Readings

Eco, U. (1990). *The limits of interpretation*. Bloomington: Indiana University Press.

McGuigan, J. (1992). *Cultural populism*. New York: Routledge.

McGuigan, J. (1999). *Modernity and postmodern culture*. Berkshire, UK: Open University Press.

EPSEM Sample

Sampling involves the selection of a portion of the population being studied. In probability sampling, each element in the population has a known, nonzero chance of being selected through the use of a random selection procedure. *EPSEM* refers to an *equal probability of selection method*. It is not a specific sampling method such as systematic sampling, stratified sampling, or multi-stage sampling. Rather it refers to the application of a sampling technique that results in the population elements having equal probabilities of being included in the sample. EPSEM samples are self-weighting; that is, the reciprocal of the probability of selection of each element in the selected sample is the same. Thus the base sampling weighting for each selected element in the sample is a constant equal to or greater than one (1.00).

The most common examples of equal probability of selection methods are (a) simple random sampling, (b) unrestricted random sampling, (c) systematic random sampling, (d) stratified sampling, and (e) proportionate stratified sampling. *Simple random sampling* refers to equal probability of selection element sample without replacement. *Unrestricted random sampling* refers to equal probability of selection element sample with replacement. *Systematic random sampling* refers to the selection of elements using a sampling interval and a random start. *Stratified sampling* refers to the formation of mutually exclusive and exhaustive groupings of elements. *Proportionate stratified sampling* then entails selecting a sample from the strata so that the proportion of the total sample allocated to each stratum equals the proportion of the total elements in the population in each stratum. So for example, if a stratum contains 25% of the population elements, 25% of the sample would be selected from that stratum.

A multi-stage sample can also be an EPSEM sample. The simplest example is a multi-stage design based on equal probabilities of selection at each stage of sampling. A more common practical example is a multi-stage design that results in an overall equal probability of selection of each element in the population, but at each stage the probabilities of selection are not equal. In two-stage sampling the clusters are usually of unequal size (i.e., the number of elements in the clusters vary from cluster to cluster, with some small clusters and some large clusters). If a probability proportional to size (PPS) sample of clusters is drawn, the larger the cluster the greater its probability for selection. So, at the first stage, the probabilities of selection are unequal. At the second stage of sampling, an equal number of elements are selected using simple random sampling from the sample clusters. So, at the second stage, the within-cluster probability of selection of an element is higher if the cluster is smaller. However, the product of the first-stage selection probability of the cluster and the second-stage probability of selection of the element within the cluster is a constant for all elements in the population. Thus, an EPSEM sample is achieved.

Michael P. Battaglia

See also Multi-Stage Sample; Hansen, Morris; Probability Proportional to Size (PPS) Sampling; Probability Sample; Simple Random Sample; Systematic Sampling; Stratified Sampling

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Equal Probability of Selection

Survey samples can be chosen in many ways, and one common approach is to use a technique that provides an equal chance of selection to all elements in the sampling frame. One type of equal probability sample is a simple random sample, but there are many others.

Morris H. Hansen, William N. Hurwitz, and William G. Madow appear to have been the first to refer to them as *EPSEM samples* ("equal probability selection method"), but the term was so often used by Leslie Kish that some have misattributed the coinage to him. Others have used the phrase *self-weighting sample*, although some eschew this term, given that weighting typically involves nonresponse adjustment and some form of calibration such as ratio adjustment or raking, and these lead to unequal weights even when all elements of the sample have been selected with equal probability. Typically, the *equal* in the title refers only to marginal inclusion probabilities. Joint probabilities of selection vary across pairs of units for designs other than simple random samples.

The variation across pairs of units is caused most often by systematic selection, stratification, clustering, or some combination of these, although it can also be caused by other sampling systems, such as controlled selection and maximization (or minimization) of overlap with other samples. The purpose of varying the joint probabilities of selection is to improve efficiency by exploiting auxiliary information. The reasons to keep the marginal inclusion probabilities constant are less compelling and largely involve tradition.

One of the innovations that was introduced in the 1940s at the U.S. Census Bureau is a scheme for multi-stage sampling that preserves equal probabilities and is very efficient. In this design, clusters are stratified into strata that, in addition to being internally homogenous, are nearly equal in population. Two clusters are then selected with probability proportional to population from each stratum. Within sample clusters, second-stage probabilities of selection are calculated so as to achieve an EPSEM sample. Given reasonably accurate population measures, this procedure will result in nearly equal-sized cluster workloads, convenient for a local interviewer to handle. Attendant reductions in the variation in cluster sample size and in sampling weights also improve efficiency.

Also, in the 1940s, it was much harder to deal with unequal weights at the analysis phase. Now, with software like SUDAAN, WesVar, and various SAS procedures that are readily available and designed to cope with unequal weights, there is less reason to design EPSEM samples. There are, however, still some reasons to consider them. Some are articulated by advocates of inverse sampling, a procedure whereby an EPSEM sample is extracted from a larger sample. Certainly, if one is interested in multi-level modeling, then an EPSEM sample can still be advantageous because there is considerable debate about how to use sampling weights in fitting such models. Another advantage arises in the context of hot-deck item imputation. If probabilities of selection are equal, then the contentious question of whether to use the weights in donor selection is avoided.

Despite these analytic and workload advantages, samplers should feel free to vary probabilities of selection using optimal allocation when advance knowledge of strata characteristics is available. This is particularly important for oversampling of minority populations in the United States.

David Ross Judkins

See also EPSEM Sample; Hansen, Morris; Hot-Deck Imputation; Inverse Sampling; Kish, Leslie; SAS; SUDAAN; WesVar

Further Readings

Anonymous. (2000). *WesVar 4.0 User's Guide* (Appendix D). Rockville, MD: Westat.

Anonymous. (2006). *SAS OnlineDoc®* (9.1.3). Cary, NC: SAS Institute.

Haziza, D., & Rao, J. N. K. (2003). Inference for population means under unweighted imputation for missing survey data. *Survey Methodology*, *29*(1), 81–90.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, A*, *169*(4), 805–827.

Rao, J. N. K., Scott, A. J., & Benhin, E. (2003). Undoing complex survey data stratures: Some theory and applications of inverse sampling. *Survey Methodology*, *29*(2), 107–128.

Research Triangle Institute. (2004). *SUDAAN language manual*. Release 9.0. Research Triangle Park, NC: Author.

Waksberg, J., Judkins, D., & Massey, J. T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, *23*, 61–71.

Error of Nonobservation

Errors of nonobservation refer to survey errors that are related to the exclusion versus inclusion of an eligible respondent or other sample record. This term principally refers to sampling error, coverage error, and nonresponse error. This is distinguished from errors of observation, which refer to errors that are related to the measurement of the content of surveys.

The term *errors of nonobservation* is based on the language and assumptions of survey methodology. It is similar to the concepts that psychometricians use to call the errors that impact external validity and, in some respects, is similar to what economists call "selection bias."

Within the total survey error perspective, errors of nonobservation can impact both random error and systematic error. Traditionally, however, coverage error and nonresponse error have been seen as being most problematic in terms of systematic error or bias. In contrast, in probability samples, sampling error is primarily seen as impacting variability, although systematic bias can also result from nonprobability samples or from inappropriate data adjustment or weighting of data from probability samples.

Sampling Error

Inference from sample surveys assumes that an underlying population is being studied and that samples are taken from this underlying population. Sample statistics, including sampling errors, are calculated to determine the variability of a statistic as measured in a survey compared to the actual or true value of that statistic in the population.

Since not all members of a population are included in a sample, survey statistics are usually different from population values. For any population, there are all sorts of possible combinations of records that might be included in any particular sample. In many cases, the results of a survey will be close to what would be found in an underlying population; in some cases they may be far off. The sampling error is traditionally taken as a measure of how the statistics obtained from any particular survey might differ or vary from those of the actual underlying population.

In terms of understanding errors of nonobservation, sample errors from probability samples primarily refer to errors regarding certainty about how close a survey statistic comes to the actual value of the statistic in an underlying population. That is, nonobservational errors due to sampling primarily impact the variability of survey statistics or the precision of the survey measure. Although there is almost always error in the form of variance, because survey results are rarely exactly in line with population statistics, these variable errors are random and thus cancel each other out across many samples.

The characteristics of sampling error are primarily mathematical and are based on several assumptions. Sampling statistics assume that a sample of respondents or other units is taken from an underlying collection, list, or frame of all members of a population.

Sampling statistics also assume that data are collected from all selected records. Moreover, probability sampling assumes that all sampled records have a known, nonzero probability of being selected.

Nonprobability samples select respondents in ways that do not permit the understanding of the specific probability that sampled members of the population are included in the sample. Convenience samples, for example, select respondents that are easily accessible to the researcher while excluding others. These sampling methods can lead to bias, when the results of measured statistics systematically differ from population values, usually in unknown ways. Bias or systematic error can also occur in scientific samples when different sample records are selected with varying likelihoods or probabilities of selection, but this bias can be adjusted for with simple mathematical adjustments known as *sample weights*.

Coverage error and nonresponse error, in contrast, come about when the sample frame does not well match the underlying population or when data are not collected from all valid sample records. In these cases, population members are not included in a survey either because they do not have a chance of being included in samples or because eligible sampled respondents do respond to a survey for various reasons.

Coverage Error

Coverage error refers to the error that occurs when the frame or list of elements used for a sample does not correspond to the population a survey is intended to study. This can occur in several ways. For example, some sample records might correspond to multiple members of a population. In contrast, some sample records might be duplicates or correspond to the same member of a population. The most problematic situation is undercoverage, where a sample frame excludes some members of the population it is intended to cover.

The primary danger involved in coverage error is coverage bias, which can occur when a sample frame systematically differs from the population it is intended to include. The extent of coverage bias depends both on the percentage of a population that is not covered in the sample frame and the differences on any statistic between those included in the sample frame and those excluded from the sample frame. For example, household surveys systematically exclude persons who are homeless, telephone surveys systematically exclude persons who do not have telephone service, and most telephone surveys have systematically excluded people who have cellular (mobile) telephone service but not traditional residential landline telephone service. In cases where the excluded proportion of a survey's target population is small, and where differences between sampled respondents and others are small, researchers usually do not have to worry about bias in results because of these exclusions. However, if the magnitude of this coverage error is large, or if the differences between covered and noncovered respondents are great, or if a survey is attempting to make very precise estimates of the characteristics of a population, then nonignorable coverage bias may result.

Nonresponse Error

Nonresponse error refers to error that occurs when persons or other elements included in a sample fail to respond to a survey. There are two types of nonresponse to surveys: unit nonresponse and item nonresponse. *Item nonresponse* (i.e., missing data) occurs when a respondent who completes a survey fails to provide answers to a question. Many Americans, for example, refuse to tell survey researchers their income. *Unit nonresponse*, in contrast, occurs when sampled respondents fail to respond to a survey at all.

Unit nonresponse occurs for a variety of reasons. Some respondents are unable to complete a survey, for example, because of a health condition or because they speak a language other than the ones a survey is administered in. Other respondents are unavailable to complete a survey, for example, because they are not at home when an interviewer calls. Still other respondents refuse to complete surveys and thus are not included in final data.

The primary danger involving survey nonresponse is potential bias in results. Similar to coverage error, the magnitude of nonresponse error depends both on the percentage of the population who fail to respond to a given question and the differences between respondents and nonrespondents on any survey statistic.

Chase H. Harrison

See also Coverage Error; External Validity; Missing Data; Nonresponse Error; Random Error; Sampling Error;

Self-Selection Bias; Systematic Error; Total Survey Error (TSE); Undercoverage

Further Readings

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.
Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.
Groves, R. M., Dillman, D. A., & Little, R. L. (2001) *Survey nonresponse*. New York: Wiley.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

Errors of Commission

Errors of commission are sometimes also called "false positives." They refer to instances in which someone or something is erroneously included for consideration when they or it should have been excluded. In survey research, this error typically occurs when the eligibility of a unit is determined. For example, when someone is screened at the start of interviewer contact to determine whether or not he or she is eligible, and the person answers erroneously in a way that makes the interview proceed as though the person were eligible for data collection when in fact the person is not eligible, then this is an error of commission. In these cases, data are gathered and analyzed for someone who should not have been interviewed.

Errors of commission occur for many reasons, but most are due to questionnaire-related error, interviewer-related error, and/or respondent-related error. The introduction to a questionnaire, where eligibility screening is typically carried out, might be worded poorly and thus cause incorrect data to be gathered, leading to people who in fact are not eligible being erroneously treated as eligible. An interviewer may administer an eligibility screening sequence poorly, thus causing the respondent to misunderstand what is being asked, leading to answers that result in an ineligible person being treated as eligible. Finally, a respondent may be unable to understand the eligibility screening questions and thus may give incorrect answers. Or the respondent may not be paying enough attention to the screening questions or may not be willing or able to give an accurate answer. In fact, some respondents will give an answer they believe will disqualify them even if it is not true for them, whereas others will do the opposite just so they can take part in the survey.

An example of errors of commission routinely occurs when people in the United States are sampled via telephone by the Nielsen Company to ask them to participate in a week-long TV diary survey. In the age of time shifting, when many people have digital video recorders (DVRs) that allow them to shift the time they view television programming, a special diary has been devised to measure programming that is time-shifted. This diary places much more of a cognitive burden on respondents who receive it than does the regular TV diary that does not measure time shifting. Thus, it would be ideal if only those people who have DVRs were sent the special diary. However, despite extensive R&D testing of the eligibility questions that are used to determine whether or not someone has a DVR, many people (especially older adults) appear not to be able to accurately answer the questions and therefore incorrectly are sent the special diary when they should receive the simpler diary.

Another example of errors of commission concerns the current situation in the United States when interviewing people reached via cell phone. Currently there are federal regulations affecting number portability that allow people to take their cell phones to another state without changing their numbers when they move. If, for example, a telephone survey of residents of New York were to be conducted without adequate screening to determine whether or not the person reached was in fact a New York resident, then all those people who had cell phones with New York area codes but now were living in other states would incorrectly be treated as eligible.

How to reduce the chances that errors of commission will occur is an important issue for researchers to think carefully about. It is not likely that such errors can be eliminated completely from surveys, and sometimes it is very expensive to institute the procedures that will keep these errors to a minimum.

Paul J. Lavrakas

See also Coverage; Coverage Error; Eligibility; Errors of Omission; Interviewer-Related Error; Number Portability; Questionnaire-Related Error; Respondent-Related Error;

Screening; Unit Coverage; Within-Unit Coverage; Within-Unit Coverage Error

Errors of Omission

Errors of omission are also sometimes called "false negatives." They refer to instances in which someone or something is erroneously excluded from consideration when they or it should have been included. In survey research, this error typically occurs when the eligibility of a unit is determined. For example, when someone is screened at the start of interviewer contact to determine whether or not he or she is eligible, and the person answers erroneously in a way that keeps the interview from proceeding as though the person were ineligible for data collection when in fact the person is eligible, this is an error of omission. In these cases, data are not gathered from someone who should have been interviewed.

Errors of omission occur for many reasons, but most are due to questionnaire-related error, interviewer-related error, and/or respondent-related error. The introduction to a questionnaire, where eligibility screening is typically carried out, might be worded poorly and thus cause incorrect data to be gathered, leading to people who in fact are eligible being erroneously treated as ineligible. An interviewer may administer an eligibility screening sequence poorly, thus causing the respondent to misunderstand what is being asked, thus leading to answers that result in an eligible person being treated as ineligible. Finally, a respondent may be unable to understand the eligibility screening questions and thus may give incorrect answers. Or the respondent may not be paying enough attention to the screening questions or may not be willing or able to give an accurate answer.

An example of an error of omission occurs whenever a survey screening sequence is worded in such a way that it is readily apparent to the respondent what answers will qualify her or him for the interview and what answers will disqualify her or him. Since many people are reluctant to directly refuse to participate in surveys in which they have little or no interest, such screening sequences are an easy way for the respondent to get out of doing the survey without outright refusing the interviewer. Thus, it is very important for researchers to structure their eligibility screening questions in ways that do not telegraph to the respondents what answers will make them eligible or ineligible.

How to reduce the chances that errors of omission will occur is an important issue for researchers to think carefully about. It is not likely that such errors can be eliminated completely from surveys, and sometimes it is very expensive to institute the procedures that will keep these errors to a minimum.

Paul J. Lavrakas

See also Coverage; Coverage Error; Eligibility; Errors of Commission; Errors of Nonobservation; Interviewer-Related Error; Questionnaire-Related Error; Respondent-Related Error; Unit Coverage; Within-Unit Coverage; Within-Unit Coverage Error

Establishment Survey

An establishment survey is a survey that seeks to measure the behavior, structure, or output of organizations rather than individuals. Establishment surveys include surveys of business that are critical to our understanding of trends in the economy, such as the Economic Census conducted by the U.S. Census Bureau. However, establishment surveys also include surveys of universities and colleges, hospitals, and nursing homes. There has been considerable discussion about the best practices involved with conducting establishment surveys in recent years, as the response rates achieved by many establishment surveys have declined similar to those of household surveys. This reduction in response rates has spurred the development of a more robust literature on conducting establishment surveys, as well as investigation of how to increase cooperation through improved questionnaire design and contacting procedures. Understanding establishment surveys requires examining the ways in which they are different from household surveys by focusing on the unique sampling, survey, and questionnaire design issues that need to be considered when studying establishments, as well as effective strategies for contacting the appropriate respondents within establishments to complete these surveys.

Differences Between Establishment and Household Surveys

Establishment surveys differ from household surveys in a number of notable ways, creating some unique challenges and considerations for survey researchers who are conducting establishment surveys. The most fundamental difference between establishment and household surveys is the unit of analysis. Whereas in a household survey, the unit of analysis is the household, family unit, or an individual, the unit of analysis for an establishment survey may be either an establishment or an enterprise. An establishment could be a business operating in a particular location, a business entity reporting unemployment insurance claims, a hospital or clinic location, and so on. An example of an enterprise would be a corporation that includes one or more locations or establishments. The term *establishment survey* is used as a generic or umbrella term to refer to surveys that collect data at either the establishment or enterprise level.

A second important distinction between establishment and household surveys is the use of informants. Whereas informants to a household survey are generally selected as part of the sampling design, informants to establishment surveys are often the person or people within an organization who are the most knowledgeable about the subject matter of the survey or those who have access to the required information. Also, although household surveys generally rely on one person to answer questions about the household overall or others living in the household, there may be multiple informants to an establishment survey, depending on the survey topics and the complexity of the establishment.

The complexity of selecting proper informants to respond to establishment surveys is usually correlated with the complexity of the sampled organizations. An example of an establishment with a simple structure for identifying a proper informant would be a nursing home, where the administrator of the nursing home would generally be the desired informant. Establishments with more complex structures, such as universities or larger businesses, may have a more decentralized structure that requires additional effort to determine who is the most appropriate informant and may include multiple informants. Surveys of larger or more complex establishments conducted via telephone may require additional time to identify appropriate informants; this is especially true of establishments for which there is no contact name provided from the sample frame. For this reason, some survey firms have invested in the additional effort of pre-contacting the establishment to obtain the name of a contact person (and potential informant) prior to the full survey being fielded. Given that the informants completing establishment surveys are answering questions about the organization as a whole, or about others within the establishment, survey questions for the establishment that ask for hard, factual data about the operations of the establishment tend to be more appropriate than questions that ask opinions or require subjective evaluation.

A further difference between establishment and household surveys is related to respondent burden, which is measured differently in establishment surveys than in surveys of household or individuals. The burden of completing a household or individual survey is generally limited to the number of questions asked of a respondent (although there are some examples of household surveys requiring respondents to obtain records prior to the interview, this is not the norm), and the questions are usually designed to be answered by recall. In contrast, an establishment survey often will have some questions that require more time to answer because of the need to compile the required information. Researchers also must consider that those responding to an establishment survey most likely are doing so during working hours, which may be discouraged by establishments because it takes time away from work tasks. This is true of both mandatory and voluntary surveys, although completion of government mandatory surveys that establishments regularly complete tends to become one of the regular work functions of certain staff members. Due to the burden issues inherent in many establishment surveys, limiting the number of survey topics involved with completing the questionnaire is an important factor in achieving high response rates.

Establishment Survey Design

Like all surveys, the design of establishment surveys requires careful consideration of a number of issues to ensure unbiased estimates of the population being studied. Several of these issues warrant discussion, as they have implications that are specific to establishment surveys, including sampling, survey design and development tools, questionnaire design, choice of mode, and effective data collection techniques.

Sampling

Establishment survey samples are generally drawn from list frames, such as business registers, or lists of other types of organizations created from multiple data sources. The sample design of many establishment surveys uses probability proportionate to size (PPS) sampling or simple random sampling with some stratification. Surveys using stratification only also tend to use the size of the establishment, since the size of the establishment generally contributes much of the variance in these list samples. For example, large surveys of establishments collecting employment data have found that employment is skewed heavily toward larger establishments, meaning that the sampling design must account for this to obtain an accurate estimate. Establishment surveys often tend to employ a second sampling stage in which those within the establishment (employees, patients, etc.) are sampled within the establishment unit.

Survey Design Issues and Development Tools

Surveying establishments requires a thorough understanding of the workings of the organizations being studied, so that the limitations on the types of data that can be collected are understood and so that the most successful approaches to obtaining cooperation can be employed. During the design phase of an establishment survey, it is useful for researchers to understand the types of data the sampled establishments are able to provide, and in what format these data tend to be kept by establishments. Similar to household surveys, the terms and definitions used in questionnaires, instructions, and other materials related to the questionnaire should be reviewed carefully to ensure they are understood in the same way by different establishments in the population. Terms that are confusing or can be understood in different ways by different respondents tend to increase item nonresponse and increase response bias by encouraging respondents to estimate answers to survey questions in ways they would not if the terms and definitions were clear. Qualitative tools such as expert panels, site visits, focus groups, and cognitive interviews, that have been used extensively in household surveys, have become more common in designing and improving establishment survey questionnaires.

Establishment surveys often rely on expert panels or stakeholders' meetings to review proposed survey items and to help researchers use language and terminology in the questionnaire and other materials that will be relevant and appropriate. This is especially important for establishment surveys of populations that use very specific terminology (such as medical practices), so that the language used in communication with sample members mirrors that used among the establishment professionals themselves.

The design of many establishment surveys has gone beyond expert review to include site visits to potential establishment respondents, and these site visits can provide researchers valuable information about the survey response process. Specifically, conducting site visits can shed light on a number of issues related to the questionnaire design process, including how establishments in the population fulfill requests to complete surveys and who in the organization handles the requests. Another potential discussion topic for site visits is the recordkeeping system or systems typically employed by establishments in the survey population, because knowing how information is recorded and stored at the establishments will help researchers to align the data requests with what establishments can provide with the highest quality and smallest possible burden. Finally, since establishment surveys often collect data about specific time periods, it is important to understand how any "reference" dates the survey plans to use mesh with how establishments in the sample receive, update, and store those data. Site visits could include unstructured one-on-one interviews, focus groups, or in-depth discussions with potential survey informants about the barriers to responding to a survey of their establishment. Surveys using these techniques prior to developing the questionnaire tend to get a solid grounding in how requests for certain data in a questionnaire will be received, whether the definitions used are in common usage and are understood in a consistent way, and the burden involved in assembling and reporting the data.

Questionnaire Design and Pretesting

There are questionnaire design issues that are particular to establishment surveys. With surveys of all types, shorter questionnaires generally produce higher response rates. Given the amount of time required by the establishment to collect information to respond to some surveys, limiting the number of data items will increase the likelihood of response by the sampled establishments. There has been increased attention to the appearance of establishment survey questionnaires in recent years, and methodological studies of the effects of layout and other visual cues have improved the usability of many establishment surveys. Improving the layout and visual presentation of establishment surveys has been shown to increase the likelihood of response by reducing the perceived burden of completing the questionnaire. Due to the number and complexity of the terminology used in the instructions, some establishment surveys have placed sets of complex definitions and instructions in a separate document to avoid the cluttered appearance that placing these items in the questionnaire itself can create.

Similar to household surveys, pretesting establishment survey instruments prior to fielding can identify problems with comprehension and survey flow and can also validate improvements made to the instrument as a result of previous design steps. Also, the use of cognitive testing has increased in recent years as researchers have realized the extent to which terms and definitions can be understood differently by respondents. Methodological work on cognitive interviewing on establishment surveys has shown how both concurrent and retrospective think-aloud techniques can be used to improve questionnaires in establishment surveys.

Mode Considerations

The modes that are used in establishment surveys have evolved considerably in recent years. For many years, mail was the predominant mode used when surveying businesses and other institutions due to the low cost of mail surveys. Similar to recent trends in household surveys, surveys offering sampled establishments a choice of modes are now common. The increasing popularity of offering a Web survey response option to selected establishments is primarily due to the fact that Web access in workplaces is now widespread. Also, the use of combinations of mail, Web, and other modes such as touchtone data entry in a survey reduces mode effects, since these modes are self-administered and questions can be displayed in essentially the same format. Surveys using mail and Web have also employed telephone contacts with establishments as reminder prompts or to collect the data as a last resort. In contrast, for voluntary surveys or surveys among populations that have shown themselves to be reluctant to participate in surveys in the past, the combination of a Web and telephone mode offering is often necessary so that interviewers can personally persuade sample members about the benefits of participating in the survey, communicate effectively with gatekeepers, and allay concerns about confidentiality.

Data Collection

As discussed previously, the level of burden associated with completing establishment surveys has depressed response rates in recent years, and maintaining similar response rate levels on periodic surveys has been accomplished only through more intense follow-up efforts and the employment of more effective data collection techniques. One of the techniques employed in establishment surveys is ensuring that the initial contact to respondents in establishment surveys is personalized, professional, and succinct. Personalizing an advance mailing by sending it to an individual rather than to the establishment at large increases the probability that the mailing will be opened, and if the targeted contact is not the most appropriate person to respond to the survey, he or she is more likely to pass along the materials to someone else at the establishment who is better able to provide the information. However, efforts to direct materials to specific individuals at the establishment are dependent on the information available from the sampling frame or on prefield efforts to identify appropriate respondents.

The design of a successful establishment survey includes a thorough understanding of the terminology used among the establishments studied, as discussed earlier. In the same vein, successful contacting of establishments during data collection should be done by staff who are familiar with this terminology and can provide relevant information about the value of completing the survey. Organizations conducting establishment surveys that include telephone interviewing or prompting have found it useful to provide additional specialized training to interviewing staff in order to avoid refusals and increase survey response. Also, a highly trained and knowledgeable staff will be better able to negotiate with gatekeepers, who tend to be more prevalent in establishment surveys due to the nature of organizations and the need for decision makers to insulate themselves from communication not central to their organizations' missions.

Use of Incentives

Similar to household surveys, the use of incentives in establishment surveys has increased, and it is generally considered necessary to provide incentives to populations that are studied regularly or that typically have low participation rates, such as physicians and business executives. Incentives for respondents at establishments tend to be more effective in fostering participation when they are explained as being "tokens of appreciation" rather than payment for the respondent's time, since busy (and highly compensated) respondents may not see the incentive as equitable value for their time spent responding.

Longitudinal Establishment Surveys

Many establishment surveys are conducted regularly to enable time-series data analyses, and often the same set of businesses is asked to complete surveys at regular intervals, such as annually. These longitudinal surveys face additional challenges, such as respondent fatigue, and the survey instruments used are generally changed very little between rounds of the survey in order to maintain time-series continuity. Even though questions on these surveys remain fairly static over time, pretesting (including cognitive interviews prior to data collection or response behavior follow-up surveys after respondents complete the survey) can identify areas of confusion that respondents may encounter while completing the survey. The decision to change survey items on longitudinal surveys should factor in the potential impact on respondents, some of whom complete the survey in multiple years and may be frustrated by changes in the measures or definitions. That concern must be balanced with the knowledge that there will also be respondents completing the survey for the first time, either due to staff turnover or because of new establishments added to the sample. Longitudinal surveys should strive to accommodate both situations to ensure similar response rates between establishments that are new to the sample and those completing the survey multiple times.

David DesRoches

See also Advance Contact; Cognitive Aspects of Survey Methodology (CASM); Cognitive Interviewing; Directory Sampling; Gatekeeper; Informant; Longitudinal Studies; Probability Proportional to Size (PPS) Sampling; Questionnaire Design; Respondent Burden; Sampling Frame; Touchtone Data Entry

Further Readings

Cox, B. G., Nanjamma, B., Kott, P. S., Colledge, M., Binder, D. A., & Christianson, A. (Eds.). (1995). *Business survey methods*. New York: Wiley.

Clayton, R. L., Searson, M. A., & Manning, C. D. (2001). Electronic data collection in selected BLS establishment programs. *Proceedings of the Second International Conference on Establishment Surveys*.

Dillman, D. A. (2001). Procedures for conducting government-sponsored establishment surveys: Comparisons of the total design method (TDM), a traditional cost-compensation model, and tailored design. *Proceedings of the Second International Conference on Establishment Surveys*.

Phipps, P. A., Butani, S. J., & Chun, Y. I. (1995). Research on establishment-survey design. *Journal of Business & Economic Statistics*, *13*(3), 337–346.

Goldenberg, K. L. (1996, May). *Using cognitive testing in the design of a business survey questionnaire*. Presented at the 51st annual conference of the American Association for Public Opinion Research, Salt Lake City, UT.

Zarkin, G. A., Garfinkel, S. A., Potter, F. J., & McNeill, J. J. (1995). Employment-based health insurance: Implications of the sampling unit for policy analysis. *Inquiry*, *32*, 310–319.

Ethical Principles

In the discipline of survey research, *ethical principles* are defined as the standard practices for privacy and confidentiality protection for human subject participants. Ethical principles in survey research are in place to protect individual participant(s) beginning at the start of study recruitment, through participation and data collection, to dissemination of research findings in a manner that is confidential, private, and respectful. These principles guide accepted research practices as they apply to the conduct of both quantitative and qualitative methods in survey research.

Background and Historical Perspective

The need for adequate protection of the individual participant and the adoption of stricter ethical principles in practice has been shaped by several historical events. One of the worst cases of ethical misconduct by researchers in the United States was the Tuskegee Syphilis Study (1932–1972) conducted by the U.S. Public Health Service with 399 black men in Macon County, Alabama, to study the natural history of syphilis. Following incidents of research malpractice such as the Tuskegee Study and the increased concerns of the public, the Nuremberg Code, the Belmont Report, and the Declaration of Helsinki were adopted.

The *Nuremberg Code* declares the need for voluntary consent by human subjects, disclosure of study procedures, and protection for vulnerable populations. The *Belmont Report*, released in 1979 by the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, emphasizes three basic ethical principles to guide practice: (1) respect for persons, (2) beneficence, and (3) justice. The principle of respect for persons emphasizes that researchers acknowledge autonomy (an individual's ability to independently make decisions) and the protection of those with diminished or restricted autonomy (individuals who are not capable of self-determination) to enter into research participation informed and voluntarily. Beneficence exemplifies the ethical treatment of persons by securing their protection from harm or risk involved with participation and the disclosure of benefits associated with study involvement. The third principle, justice, encourages equal treatment and provision of advantages and access associated with research to all persons. The *Declaration of Helsinki*, adopted in 1964, was designed to guide physicians in biomedical research involving human subjects to safeguard the health of people.

Respondent/Participant

Informed Consent

Informed consent is designed to protect survey participants' rights to voluntary participation and confidentiality and thus relates to basic concerns over respect, beneficence, and justice, as discussed previously. To make an informed decision to participate, individuals must understand that the survey involves research and the purpose of the research. The consent statement should communicate the expected burden (typically for survey research, the length of commitment) and any potential discomfort that may result from participation, such as distress resulting from the sensitivity of questions. The consent statement likewise should describe potential benefits that the survey may have on the individual and society. Informed consent requires an explanation of the voluntary and confidential nature of the study.

The consent statement must be understandable to potential study participants and thus avoid jargon and complex language. In some survey situations, study participants may be asked to agree to several activities, such as answering questions in a survey and being audiotaped. Each individual activity should be presented on its own and consent obtained separately.

Voluntary Participation

Participation in survey research is fully voluntary, meaning that individuals have the right to decide themselves whether they wish to participate. Once that decision is made, participants have the right to withdraw from the survey at any time or refuse to answer any question. The voluntary nature of participation requires that such decisions be made without negative consequence or withdrawal of benefits to the participant.

Confidentiality and Privacy

Survey participants have the right to control the disclosure of information about themselves that they may consider private. Information that is considered private may vary across communities and individuals. Survey researchers have the responsibility to treat any information shared in an interview setting or on a self-administered questionnaire as confidential. All individuals involved in collecting data should be fully trained in confidentiality practices and be required to adhere fully to such protocols without exception.

The design of the survey instrument itself affects the protection of privacy. At times, surveys may ask about sensitive topics that participants may view as intrusive to their privacy. The consent form should clearly divulge the topics covered in the interview, and the survey design itself should allow participants to choose not to answer any question. The storage of survey data, including audio- and videotapes, clearly impacts the process of maintaining confidentiality. The first consideration is whether participant identifiers need to be kept. If so, survey practitioners should consider options for protocols that limit access and prevent inadvertent disclosure.

In some situations, private data can be protected from subpoena by third parties. Studies that collect information potentially harmful to individuals if disclosed, such as substance abuse, other illegal behaviors, or sexual practices, may be granted a *Certificate of Confidentiality* issued by the National Institutes of Health that offers protection against disclosure. Any study, regardless of funding or source of funding, can apply for a Certificate of Confidentiality. Limitations to confidentiality in a survey study do exist. For example, some states require the reporting of child abuse and neglect or intent to harm oneself or others. In survey situations that may uncover such events, the interviewer may be required to report this to her or his supervisor, who in turn would follow the survey organization's procedure for such cases.

Incentives

Incentives in survey research are often used as a vehicle for thanking study participants for their contributions or as a means for initially motivating interest and participation. Incentives may pose ethical conflicts when their appeal is too strong (i.e., coercive) and thus misrepresent the voluntary nature of the study or bias the participants' decision to participate.

Vulnerable Populations

Some study populations may be more vulnerable to violation of their rights to informed consent and voluntary participation. Such populations include minors, prisoners, cognitively disabled individuals, economically and educationally disadvantaged individuals, and those with mental health conditions. Survey researchers have responsibility to build in additional protections for these individuals. For example, such protections may include building in an oral consent requirement when literacy levels are low in a study population or, in the case of research with children, obtaining parental consent.

Researcher/Investigator

Ethical principles also provide a foundation to guide the researcher in the design of a study. The principles supply a standard for practice that is used to guide research design, conduct, analysis, and reporting of findings. Applied and ethical theories are the basis for several codes of professional ethics and practices (e.g., American Association for Public Opinion Research [AAPOR] and Council of American Survey Research Organizations [CASRO]) that are based on teleological, deontological, and casuistical ethics. Teleological ethics in professional practice guide the practice and development of research studies that aim to answer a set of questions in a manner that is valid and reliable with an outcome that is meaningful for a population. Deontological ethics are those that are inherently and morally correct (protection of confidentiality, voluntary participation). The codes of professional ethics and practices guiding research are established on accepted principles that model best practices in the field of survey research and are casuistical.

Research Design, Practice, Reporting, and Disclosure

Survey researchers should aim to design research that protects participants through methods and practices that are approved by institutional review boards (IRB), or their equivalent in the private sector, and comply with ethical principles. Survey researchers should be committed to the conduct of research of the highest integrity that carefully weighs and accounts for participant benefits, risk, and protection. Ethical principles extend beyond recruitment and data collection to the management of data, analysis, and the report of findings. Analysis of data collected should not include identifying variables that could lead to a violation of survey participant confidentiality.

Survey practitioners have ethical obligations to the public to share their methods and findings. Disclosure of the study design and methods furthers the goals and advancement of science generally, but more specifically, allows for others to reflect carefully on the strengths and limitations of the study itself and allows for replication of research and validation of findings.

The findings released from a survey should accurately reflect the survey data. Misrepresentation of survey results can violate society members' rights to respect and justice, and in some situations result in harm to community standing, access to services and programs, and personal freedoms.

Researchers are required to report study methods that include procedures of participant recruitment, data collection (participant and nonparticipant characteristics), data management (editing and coding), and analysis that allow for adequate peer review, replication and validation of research findings, and the evaluation of the quality of published studies. Full disclosure of a survey and its findings should include several key pieces of information, such as (a) research sponsorship and name of the organization that conducted the survey; (b) a description of survey objectives; (c) sampling frame and sampling methodology; (d) dates of the study; and (e) exact wording of the questions. The reporting of survey methods in a manner that protects participant confidentiality yet allows for professional evaluation of the quality of survey research avoids the creation of publication bias and misinterpretation and misleading conclusions. Survey researchers are expected to avoid falsification of findings and plagiarism.

As new methods of data collection (e.g., Web surveys) are introduced and accepted into practice, ethical principles must be modified and implemented to assure the protection of participants. Analysis techniques (e.g., Hierarchical Linear Modeling and Geographic Information Systems) continue to evolve and allow survey researchers to report and illustrate findings in ways that assure participant confidentiality and informed consent.

Through practice, survey researchers have learned that ethical principles not only protect individual rights in research participation but contribute to better research methods that may increase the trust between survey participants and research interviewers and result in better rapport and quality of data collected. Ethical principles have also improved confidence in the value of research within specific communities as well as overall public trust.

Melissa A. Valerio and Tina Mainieri

See also American Association for Public Opinion Research (AAPOR); Beneficence; Certificate of Confidentiality; Confidentiality; Consent Form; Council of American Survey Research Organizations (CASRO); Disclosure; Informed Consent; Institutional Review Board (IRB); Voluntary Participation

Further Readings

American Association of Public Opinion Research: http://www.aapor.org/aaporcodeofethics?s=aapor%20code
American Psychology Association: http://www.apa.org/ethics/code2002.html
American Sociological Association: http://www.asanet.org
American Statistical Association Privacy, Confidentiality, and Data Security: http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main
The Belmont Report: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm
Council of Survey Research Organizations Code of Standards and Ethics for Survey Research: http://www.casro.org/codeofstandards.cfm
Declaration of Helsinki: http://www.cirp.org/library/ethics/helsinki
Office for Human Research Protections (OHRP), National Institutes of Health: http://www.hhs.gov/ohrp
Qualitative Research Consultants Association: http://www.qrca.org/displaycommon.cfm?an=1&subarticlenbr=26

EVENT HISTORY CALENDAR

The event history calendar is a conversational interviewing approach that is designed to collect retrospective reports of events and the timing of their occurrences for reference periods that can range from several months to an entire life course. Different researchers have used different terms, including *life history calendar, life events calendar, crime calendar, life events matrix*, and *neighborhood history calendar*.

Key Components

The primary aim of the event history calendar approach is to maximize the accuracy of autobiographical recall. Just as event history calendars represent the past both thematically and temporally, the structure of autobiographical knowledge is believed to be organized in a similar fashion. Theoretically, the thematic and temporal associations of events within the structure of autobiographical knowledge afford retrieval cues that can be implemented in event history calendar interviewing and aid respondents to reconstruct their pasts more completely and accurately. One type of retrieval cue involves the sequencing of periods of stability and the transitions between them with regard to what happened earlier and later in time within the same timeline. For example, one may remember that one's employment period in one company immediately preceded another period of employment with a different company. In between these periods resides the transition point from one period to another, and both the length of periods and the timing of transition points are recorded within event history calendar timelines. In addition to sequential retrieval, the use of parallel retrieval cues involves the remembering of events across timelines and domains that happened contemporaneously, or nearly so. For example, one may remember a period of unemployment that was contemporaneous with a change in residence from one location to another. Parallel retrieval is particularly effective if the timing of one of the events is especially memorable, as this memory will locate the timing of the other event as well. It is the use of the respondents' own remembered events as cues to recall less easily retrieved information that is hypothesized to lead to benefits in data quality, and it is this requirement that necessitates conversational flexibility in interviewing.

Other key components of event history calendar design include representing thematic aspects of the past into domains, such as residence and labor, and the capturing of temporal changes by the inclusion of one to several timelines within each domain. For example, a labor history domain may include separate timelines to collect temporal changes in the amount of work, periods of unemployment, and periods in which one had been out of the labor force. In addition to the length of the reference period, instrument designers need to determine the smallest units of time in which life events are to be located, whether years, months, or in some instances, thirds of a month. For longer reference periods, larger time units should be used, so as to provide participants with a level of temporal detail that is best matched to how finely tuned the timing of events can be reconstructed in memory and to lend to both interviewers and respondents a manageable number of units. The first domains that should be queried are those whose events are most easily remembered, to motivate responding and also to lay out a framework in which more easily remembered events can be used as cues in the remembering of events that are queried later in the interview. Requesting respondents to provide "landmark events," such as the timing of holidays and birthdays, can be an effective first domain when used in this fashion, but landmarks appear most beneficial for shorter reference periods. With longer reference periods, it may be best to ask respondents to trace their residential histories, which helps respondents to map temporal locations with physical ones. As part of their use of flexible conversational interviewing, event history calendars can be implemented by allowing interviewers and respondents to return to domains once covered, although usually the interview flows by proceeding from one domain to the next.

Background

The first administrations of the event history calendar interviewing methodology can be traced to 1969, with the Monterrey Mobility Study of 1,640 men ages 21–60 in Monterrey, Mexico, and with the Hopkins Study that recruited a U.S. national probability sample of 953 men ages 30–39. The method became more widely recognized as a viable approach to questionnaire design in the late 1980s, with its implementation in the University of Michigan's Study of American Families. Complementing these earliest efforts in the fields of demography and sociology, the event history calendar has been administered by researchers in a variety of disciplines, including criminology, economics, nursing, psychiatry, psychology, epidemiology, social work, and survey methodology. It has been successfully used to collect mobility, labor, wealth, partnering, parenting, crime, violence, health, and health risk histories. Although aptly used in scientific investigations seeking to uncover the causes and consequences that govern well-being within populations, it has also been used in more qualitative efforts, including those in clinical settings that examine an individual's patterns of behavior to assess potential beneficial interventions.

Event History Calendar Versus Standardized Interviews

Several studies have directly compared the quality of retrospective reports engendered in event history calendar and standardized interviews in experimental and quasi-experimental designs. In most, but not all, instances, the event history calendar has been shown to lead to more accurate retrospective reports. Some of the differences in data quality between interviewing methods have been impressive. For example, in reports of intimate partner violence, the event history calendar has been shown to be effective in eliminating an age-cohort bias in reports of first exposure to violence that is observed in standardized interviews. Whereas standardized interviewing leads to older women reporting the first exposure to intimate partner violence at older ages due to the failure to report earlier instances, such an age-cohort bias is not observed in event history calendar interviews. The advantages of event history calendar interviewing have yet to be isolated as the result of the use of more effective retrieval cues, the ability of conversational flexibility to repair misunderstandings and to clarify question objectives, or some combination of both retrieval cuing and conversational interviewing. Ongoing verbal behavior coding studies that document the occurrence of different types of retrieval cues and conversational mechanisms may uncover which types of verbal behaviors produce better data quality. Such work is likely to lead to improvements in interviewer training.

That event history calendars show mostly encouraging gains in data quality in comparison to standardized interviewing indicates that it is not a panacea that will "cure" all ills associated with forgetting, and that there are also likely beneficial aspects to standardization that are not utilized in event history calendar interviews. The very few studies that have been conducted have shown that event history calendar interviewing leads to modest increases in interviewer variance in most, but not all, instances. The event history calendar also usually leads to modest increases in interviewing time, at present on the order of 0%–10% longer than standardized interviews. Interviewers show overwhelming preference for event history calendar interviewing in ease of administration. As an attempt to acquire the "best of both worlds," hybrid event history calendar and standardized interviewing instruments have also been designed.

Administration Methods

Event history calendars have been administered in a variety of methods, including as paper-and-pencil and computer-assisted interviewing instruments, and in face-to-face, telephone, and self-administered modes. The method has mostly been implemented in the interviewing of individuals, but the interviewing of collaborative groups has also been done. The computerization of event history calendars affords the automation of completeness and consistency checks. Web-based applications are also being explored.

Robert F. Belli and Mario Callegaro

See also Aided Recall; Conversational Interviewing; Diary; Interviewer Variance; Reference Period; Standardized Survey Interviewing

Further Readings

Axinn, W. G., & Pearce, L. D. (2006). *Mixed method data collection strategies*. Cambridge, UK: Cambridge University Press.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, *6*, 383–406.

Belli, R. F., Shay, W. L., & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public Opinion Quarterly*, *65*, 45–74.

Belli, R. F., Stafford, F. P., & Alwin, D. F. (in press). *Calendar and time diary methods in life course research*. Thousand Oaks, CA: Sage.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology*, *18*, 37–68.

Yoshihama, M., Gillespie, B., Hammock, A. C., Belli, R. F., & Tolman, R. M. (2005). Does the life history calendar method facilitate the recall of intimate partner violence? Comparison of two methods of data collection. *Social Work Research*, *29*, 151–163.

EXHAUSTIVE

Exhaustive is defined as a property or attribute of survey questions in which all possible responses are captured by the response options made available, either explicitly or implicitly, to a respondent. Good survey questions elicit responses that are both valid and reliable measures of the construct under study. Not only do the questions need to be clear, but the response options must also provide the respondent with clear and complete choices about where to place his or her answer. Closed-ended or forced choice questions are often used to ensure that respondents understand what a question is asking of them. In order for these question types to be useful, the response categories must be mutually exclusive and exhaustive. That is, respondents must be given all possible options, and the options cannot overlap. Consider the following question, which is frequently used in a number of different contexts.

Please describe your marital status. Are you…
Married
Divorced
Widowed
Separated
Never married

This question does not provide a response option for couples who are in committed relationships but are not married, whether by choice or because of legal barriers. For example, a woman who has been with a female partner for 5 years would be forced to choose either married or never married, neither of which accurately describes her life situation. Without a response option that reflects their life circumstances, those respondents may be less likely to complete the questionnaire, thus becoming nonrespondents. This question is easily improved by the addition of another response category:

A member of an unmarried couple

In situations in which the researcher cannot possibly identify all response options a priori, or cannot assume a single frame of reference for the subject matter, an "Other [specify]" option can be added. For example, questions about religion and race always should include an "Other [specify]" option. In the case of religion, there are too many response options to list. For race, traditional measures often do not adequately capture the variety of ways in which respondents conceptualize race. Thus, an "Other [specify]" option allows respondents to describe their race in a way that is most accurate to them.

Linda Owens

See also Closed-Ended Question; Forced Choice; Mutually Exclusive; Open-Ended Question

Further Readings

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

EXIT POLLS

Exit polls are in-person surveys in which data are gathered immediately after people have engaged in the behavior about which they are being surveyed, such as voting in an election. The survey methods that are used in exit polls apply to the measurement of a wide variety of behaviors, but in the minds of most people exit polls are most closely associated with what is done on Election Day to help project the winning candidates before the final vote tally is announced. Although political exit polling is done in many countries, it is the exit polling conducted for elections in the United States that is covered here.

How Exit Polling Is Conducted and Used in U.S. Elections

The exit polls that are conducted nationwide and in most individual states for the general election in the United States are among the largest single-day surveys that are conducted anywhere, with data from more than 100,000 respondents being gathered, processed, and analyzed within one 24-hour period.

To estimate the outcome of an election in a particular geopolitical area of the United States, which most typically is done at the state level, a stratified random sample of voting precincts within the area is selected, and at least one interviewer is sent to each of the sampled precincts. In the 2004 U.S. general election, there were 1,469 sampled precincts nationwide, and in 2006 there were 993. Those exit polls were conducted by Edition Media Research and Mitofsky International, the organizations that were hired to gather the exit poll data for their news media funders (ABC, the Associated Press [AP], CBS, CNN, Fox, and NBC). On a systematic basis, and in order to obtain a completed questionnaire, the exit poll interviewer stops (i.e., intercepts) people who just finished voting as they exit from their voting places. For example, the interviewers may do this with every 10th person who comes out of the voting place. In each sampled precinct, an average of approximately 100 voters is interviewed over the course of Election Day. Not all exiting voters who are stopped agree to complete the exit poll questionnaire, but in those cases the interviewer records basic demographic information about these refusing voters. This information is used later as part of analyses that investigate the nature of exit poll nonresponse. Interviewers at each sampled precinct telephone in the data they are gathering at three scheduled times on Election Day: mid-morning, early afternoon, and within the hour before voting ends in the precinct.

In order to gather the exit poll data, the interviewer typically hands the selected voter a questionnaire on a clipboard and asks her or him to complete it and then deposit it in a survey "ballot box." The questionnaire gathers three types of data: (1) it measures who the sampled citizen voted for in the day's key election contests (e.g., president, senator, and governor); (2) it measures various attitudes held by the voter that the news media sponsors believe will be associated with the votes cast by the sampled voter; and (3) it measures key demographic and lifestyle characteristics of the voter to further help explain why he or she voted as he or she did. All of these data are gathered via a questionnaire comprised of 20–25 questions that fit onto one piece of paper, which is printed on the front and back sides.

The survey data gathered by exit polls are used for two major purposes by the news media who sponsor them. First, exit polls are used to project the outcome of races on Election Night. However, these projections are not based solely on exit poll data. Rather, the statistical models that have been devised to help make accurate election projections utilize a variety of data, depending in part on how close is the election contest that is being projected. In a "landslide" election contest, a projection can be based with confidence on (a) the prior expectations (e.g., from preelection polls and past election history for the geopolitical area) with which the statistical models are "primed," (b) any "early voter" survey data that may have been gathered, and (c) the exit poll data. In very close elections, a confident decision cannot be made on the basis of just these three types of information, and the projection model is supplemented with additional data as actual vote counts from the sample precincts become available after voting has ended. And, then later, countywide actual vote counts are used in the models as those data become available throughout the night of the election and into the early morning of the next day. In the case of the actual vote count from the sample precincts, the projection model also takes into account the match or lack there of between the exit poll data for a precinct and the actual vote in the precinct to calculate a *bias estimate* in the exit poll data for that precinct. These bias estimates are also taken into account by the model and by the human decision makers who ultimately make the final decision about when to "call" a race in very close contests. Thus, despite what many people believe, the election projections announced on television in many cases are based very little, or even not at all, on the exit poll data.

Although it is these projections of election winners that most people think of when they think of exit polls, the second use of exit poll data is arguably more influential on the geopolitical entity and the citizenry being measured. Thus, exit polls are important, not so much because they are used to help make the projections reported by the major television networks on Election Night, but because the information they gather about the voters' demographics and attitudes toward the candidates and the campaign issues provides very powerful and important explanations about *why* the electorate voted as it did. It is only through the use of accurate exit poll data that the so-called mandate of the election can be measured and reported accurately without relying on the partisan spin that the candidates, their campaign staff, and political pundits typically try to put on the election outcome. For example, in 1980, Ronald Reagan's strategists described his sound defeat of Jimmy Carter as a "turn to the right" by American voters and as an impetus for a conservative legislative agenda for the new Congress. In contrast, 1980 exit-poll data showed there was no ideological shift among American voters. Instead, they were primarily concerned about President Carter's inability to influence the economy and settle the Iran hostage crisis, and they wanted a new president whom they hoped would do a better job in reducing inflation. As another example, in the 1998 exit polls, voters indicated that they were basing their votes for Congress on evaluations of their local candidates and not on any concerns about the allegations regarding President Bill Clinton and Monica Lewinsky contained in Kenneth Starr's report.

The Evolution of Exit Polling in the United States

Exactly when the first election exit poll was conducted is debatable. Warren J. Mitofsky, who is recognized as the father of exit polling, believed it took place in November 1967 and was conducted by CBS News for a gubernatorial election in Kentucky. In contrast, I. A. "Bud" Lewis, former pollster for *The Los Angeles Times*, believed it was in June 1964 when NBC News gathered data from California voters exiting polling places in 21 sampled precincts on the day of that state's primary election.

Starting in the late 1960s and until 1990, ABC, CBS, and NBC conducted their own exit polls for major elections. In 1990, these networks joined with CNN to create and sponsor a new entity, Voter Research and Surveys (VRS), to gather exit poll data they all would share. In 1994, a further consolidation took place, and this joint operation became the Voter News Service (VNS) when the VRS exit polling operations were merged with the News Election Service, which heretofore had been responsible for gathering actual vote counts. The four television networks and the AP, which joined as a sponsor in 1994, shared the same VNS database on election nights from 1994 through 2002, although each media organization (with the exception of the AP) used its own methods, models, and expert consultants to produce its projections of winners. Fox News joined the group as a VNS sponsor in 1996. VNS provided high-quality data through the 2000 election prior to its dissolution after the November 2002 election. In 2003, the six major sponsors of VNS contracted with Edison Media Research and Mitofsky to create the 2004 National Election Pool (NEP) service. With Mifotsky's sudden death in the summer of 2006, Edison Media Research took over sole responsibility for the 2006 exit polls and continued with that responsibility for the 2008 exit polls. Through the 2006 election, only *The Los Angeles Times* (which started its own exit polling in 1980) conducted an alternative national exit poll in presidential election years.

Exit Polling and Total Survey Error

Exit polls provide survey researchers with unique data to help understand survey accuracy. Exit polls that measure election outcomes generate survey data that generally can be validated against the actual vote count in an election within a day after the exit poll data are gathered. Thus, unlike almost all other surveys, there is a "gold standard" that allows exit pollsters to calculate how right or wrong the exit poll data were and to investigate why they were not more accurate.

From the standpoint of coverage error and sampling error, the exit polls planned and implemented by Mitofsky and his colleague, Murray Edelman, during the past 3 decades have had nothing to fault them. The sampling design is an exceptional one for the purposes to which the data are to be put. The stratified samples that are drawn in each state traditionally have covered the electorate very well, and the sample sizes have provided adequate statistical power for the uses that are made of the exit poll data, both in terms of helping to project the election outcomes and to analyze why the electorate voted as it did. However, a growing coverage threat to exit poll accuracy stems from the "early voter" phenomenon. As proportionally more and more voters in certain states (e.g., Oregon, Washington, California, Florida) choose to vote before the day of an election, the electorate that must be measured is not all available on Election Day for sampling by exit polls. And, to the extent that early voters differ in their voting behavior and motivations from those who vote on Election Day, exit pollsters must field accurate pre–Election Day surveys (e.g., telephone surveys) of these early voters to combine those data with the data gathered via exit polls on Election Day. A problem with this is that telephone surveys of early voters are not as accurate a survey methodology as are same-day in-person surveys such as the exit polling done on the day of an election.

From a nonresponse error standpoint, research reported by Mitofsky in the early 1990s and additional research reported by Daniel M. Merkle and Edelman in the late 1990s has shown that the exit poll error that exists at the level of the sampled precinct—that is, the difference between the exit poll election outcome data for a precinct and the actual vote count in that precinct—is *uncorrelated* with the exit poll response rate at the precinct level. In the exit polls conducted by VRS, VNS, and NEP, the precinct-level response rate generally falls within the 30%–70% range. What this nonresponse research has shown is that whether the response rate is on the low side or the high side is not related to the size of the bias for how people have voted in a precinct. However, because of the great need for exit poll data to remain as accurate as possible, continued investigations in the possibility of nonresponse bias are needed. This is necessary, if for no other reason than to help counter partisan claims that exit polls are biased because voters of certain political persuasions are more likely to refuse to participate when they are sampled for exit polls.

From a measurement error standpoint, a great advantage of exit polls in controlling respondent-related error is that the election behaviors and attitudes of interest are being measured at almost the exact time they occurred. That is, it is just a few minutes after a respondent has made her or his final decisions about how to vote that she or he is asked in the exit poll questionnaire to answer questions about these voting behaviors and attitudes. Thus, there is essentially no chance that an exit poll respondent will have forgotten how she or he voted, and there is little chance that her or his attitudes on key matters will have changed between the time of voting and the time of completing the exit poll questionnaire. Furthermore, because the questionnaire is self-administered and is placed into a "ballot box" by the respondent, there is little chance for social desirability to bias the answers the respondent provides, since the exit poll interviewer is unaware of any of the respondent's answers. All this notwithstanding, the effects of interviewer-related error and questionnaire-related error on the measurement accuracy of exit polls is yet to be understood as well as it needs to be, and thus continued research into these topics is needed.

Paul J. Lavrakas

See also Coverage Error; Election Night Projections; Horse Race Journalism; Mall Intercept Surveys; Measurement Error; Nonresponse Error; Pre-Election Polls; Sampling Error; Social Desirability; Stratified Sampling; Systematic Sampling; Total Survey Error (TSE)

Further Readings

Lavrakas, P. J. (2007). *Why our democracy needs accurate national exit polls*. Retrieved April 5, 2008, from http://www.publicopinionpros.com/features/2007/jan/lavrakas.asp

Lavrakas, P. J., Holley, J. K., & Miller, P. V. (1991). Public reactions to polling news during the 1988 presidential election campaign. In P. J. Lavrakas & J. K. Holley (Eds.), *Polling and presidential election coverage* (pp. 151–183). Newbury Park, CA: Sage.

Lindeman, M., & Brady, R. (2006). *Behind the controversy: A primer on U.S. presidential exit polls*. Retrieved April 5, 2008, from http://www.publicopinionpros.com/from_field/2006/jan/lindeman_1.asp

Merkle, D. M., & Edelman, M. (2000). A review of the 1996 Voter News Service exit polls from a total survey error perspective. In P. J. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 68–92). New York: Seven Bridges.

Merkle, D. M., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 243–258). New York: Wiley.

Mitofsky, W. J. (1991). A short history of exit polls. In P. J. Lavrakas & J. K. Holley (Eds.), *Polling and presidential election coverage* (pp. 83–99). Newbury Park, CA: Sage.

Mitofsky, W. J. (2006). *The future of exit polling*. Retrieved April 5, 2008, from http://www.publicopinionpros.com/op_ed/2006/jan/mitofsky.asp

Mitofsky, W. J., & Edelman, M. (1995). A review of the 1992 VRS exit polls. In P. J. Lavrakas, M. W. Traugott, & P. V. Miller (Eds.), *Presidential polls and the news media* (pp. 81–100). Boulder, CO: Westview.

Traugott, M. W., & Lavrakas, P. J. (2008). *The voters' guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Experimental Design

Experimental design is one of several forms of scientific inquiry employed to identify the cause-and-effect relation between two or more variables and to assess the magnitude of the effect(s) produced. The independent variable is the experiment or treatment applied (e.g., a social policy measure, an educational reform, different incentive amounts and types) and the dependent variable is the condition (e.g., attitude, behavior) presumed to be influenced by the treatment. In the course of the experiment it is necessary to demonstrate the existence of covariation between variables, its nonspuriousness, and to show that the cause occurred before the effect. This sort of inquiry can take the form of an artificial experiment, carried out in a laboratory scenario, or a natural experiment implemented in a real-life context, where the level of control is lower. For both cases, the literature presents several taxonomies, from which four main types are considered: (1) true or classical experimental, (2) pre-experimental, single-subject experimental, and (3) quasi-experimental. In addition, there are a number of variations of the classic experimental design as well as of the quasi-experimental design.

In a true or classic experimental design, there are at least two groups of individuals or units of analysis: the experiment group and the control group. Participants are randomly assigned to both groups. These two groups are identical except that one of them is exposed to the experiment or causal agent, and the other, the control group, is not. In many instances, a pretest and a posttest are administered to all individuals in the two groups; but the pretest is not a necessary aspect of the true experiment. If there is a significant difference between members of the two groups, it is inferred that there is a cause-and-effect link between that treatment and the outcome.

The pre-experimental design does not have a control group to be compared with the experiment group. There is a pretest and a posttest applied to the same participants. In a single-subject experimental design, there is only one participant, or a small number, that is analyzed over a period of time. In quasi-experimental designs, participants come from naturally assembled or pre-determined groups (e.g., a family, a school class, a professional category, or inhabitants of a neighborhood) and are not therefore assigned randomly to the control and treatment groups.

An experimental design has to fulfill several conditions. The variables must be measured with accuracy and precision, and the statistical test must be defined before starting the experiment. If necessary, it must be possible to repeat the experiment in order to confirm that the outcome is statistically significant and that no other factors, other than the independent variable(s) the researcher manipulates, are responsible for the outcome. In practice, however, not every experiment meets all these conditions. Internal and external validity can be affected by several factors. For example, internal validity (cause-effect relation) can be influenced by the length of the time between the pretest and the posttest, by changes in the measurement instruments, the influence of the pretest on subsequent behavior, and so on, but can be improved with matching procedures and by randomization. External validity can be enhanced by the use of a representative sample and by avoiding artificial experiment settings lacking mundane realism

Experimental design is used in biological, physical, and social sciences and can be combined with other forms of scientific inquiry (e.g., experimentation can test hypotheses contained in a formal model). However, in contrast to biology, chemistry, physics, and medicine, where experimentation is widely used, in many cases experimental design is not practicable in social science research. An exception is in psychology and behavioral branches of other social sciences, such as behavioral economics or political research on voting behavior and voter decision-making processes. Due to concerns about external validity, social factors often are too complex to be validly represented in an experiment. Furthermore, due to concerns about fundamental ethical principles, other ways of assignment to ameliorative treatments (rather than randomization) may introduce biases in the selection of participants. As a consequence, experimental design has not been a dominant mode of inquiry in the social sciences. However, in survey research, it is an ideal approach for testing the effects of methodological innovations.

Experimental design with human subjects, in both social and biological sciences, raises several ethical issues, some of which may even preclude the use of experiments. For example, a study of the long-term effects of unemployment cannot be done by assigning people randomly to an employed control group and an unemployed experiment group. It would be unethical to do so and instead the study should be done with data obtained from nonexperimental studies, such as longitudinal surveys. In other cases, the experience is ethically acceptable only if the event occurs naturally. The reluctance of human participants to be used as material for experiments is another factor that makes the use of experimentation in social science more difficult. Finally, issues of informed consent and deception in research with human participants need to be addressed even more carefully in the case of a social experimental design.

Experimental designs have a number of advantages over other forms of empirical inquiry whose primarily aim is to determine the cause-effect relation between variables. It is the best research device from the point of view of internal validity, since it can reduce, at least in theory, the partiality that exists in all other forms of inquiry based on observational data. It can provide answers about causal relations that other forms of inquiry do not, and as such it is important for the development of explanations of complex social behavior (e.g., response/nonresponse to surveys). It is also useful in those circumstances when it is necessary to validate formal models or to produce evidence to support fundamental arguments.

However, experimental design also has some weaknesses: (a) it is not suitable for all kinds of studies (e.g., longitudinal studies often are better for situations of mutual causal effects); (b) it assumes that once random assignment has been adopted it will always generate valid results; (c) it can artificially produce the expected results (this weak point can be somewhat reduced by the use of double-blind practices); (d) it may shape reality quite artificially and does not fully consider other relevant factors, assuming a particular variable (treatment, information, policy measure, etc.) as the cause of the observed differences between the two groups; and (e) it is often too expensive. In social sciences, besides these general difficulties, there are additional ethical constraints that affect this model of inquiry.

Within survey research, experimentation is commonplace when investigating questionnaire effects such as wording, ordering, and formatting. It also is used routinely to study the effects of different types and amounts of incentives and other treatments (e.g., advance letters; special introductions for interviewers) to counter survey nonresponse.

Pure or classic experimental design is, under certain conditions, the most powerful form of inquiry for the identification of causal relations. However, some of those conditions, like randomization and experimental control, make its application in social science, including survey research, more problematic than in the biological or physical sciences for technical and ethical reasons. Other forms of experimental inquiry, such as quasi-experimental design, that do not involve random selection of participants can be useful research tools for the study of causal mechanisms in a myriad of social situations in geography, planning, political science, sociology, and in other disciplines as well.

Carlos Nunes Silva

See also Control Group; External Validity; Factorial Design; Internal Validity; Random Assignment; Solomon-Four-Group Design

Further Readings

Campbell, D. T., & Stanley, J. (1973). Experimental and quasi-experimental designs for research. The American Educational Research Association (1963). In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 1–76). Chicago: Rand McNally.
Christensen, L. B. (1994). *Experimental methodology*. Boston: Allyn & Bacon.
Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches* (2nd ed.). London: Sage.
Frankfort-Nachmias, C., & Nachmias, D. (2000). *Research methods in the social sciences* (6th ed.). New York: Worth Publishers.
Hakim, C. (1992). *Research design: Strategies and choices in the design of social research*. London: Routledge.
Mitchell, M., & Jolley, J. (2004). *Research design explained* (5th ed.). Belmont, CA: Wadsworth.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

EXTERNAL VALIDITY

External validity refers to the extent to which the research findings based on a sample of individuals or objects can be generalized to the same population that the sample is taken from or to other similar populations in terms of contexts, individuals, times, and settings. Thus, external validity is generally concerned with the generalizability of research results and findings to the population that the sample has been taken from. It is a very important concept in all types of research designs (true experimental, quasi-experimental, and nonexperimental) including ones that use surveys to gather data. Therefore, assuring the external validity and the generalizability of the findings should be one of the primary goals of the survey researcher.

The language of survey research often does not include the term *external validity*. However, that concept includes what survey researchers refer to as nonresponse error and coverage error, in that each of these types of survey error is related to how well the findings from a survey can be generalized to the target population the survey purports to represent.

Threats to External Validity

Threats to external validity are the characteristics of any type of research study design, including survey research, that can influence and limit the generalizability of the research findings. These threats may include (a) sample characteristics, (b) setting characteristics, (c) temporal characteristics, (d) pretesting effects, (e) multiple-treatment interferences, (f) high dropout (attrition) rates, and (g) low response rates. Avoiding and/or controlling these threats help a survey research study increase the level its external validity and thus the generalizability of the results.

Sample Characteristics

If the results of the survey apply only to the sample, rather than to the target population from which the sample was selected, then one might question why the survey was conducted in the first place, as the results are not likely to have any value. In order to avoid this situation, the researcher should make certain that the sampling design leads to the selection of a sample that is representative of the population. This can be accomplished by using an appropriate probability sampling method (e.g., simple random sampling, systematic sampling, stratified sampling, cluster sampling, multi-stage sampling) to select a representative sample from the target population. Generally this means drawing a sample that has sample characteristics (gender, age, race, education, etc.) that closely match the parameters of the target population. Nonprobability samples generally have little or no external validity.

Setting Characteristics

This threat to external validity refers to the characteristics of the survey study's setting that may limit the generalizability of the results of the study. The major concern with this threat to the external validity is that the findings of a particular survey research study may be influenced by some unique circumstances and conditions, and if so, then the results are not generalizable to other survey research studies with different settings. The research site, specific experimental setting arrangements, intervention delivery method, and experimenter's competency level are examples of such possible setting factors that well can limit the generalizability of the results.

One of the methods that can be used for minimizing the survey research setting threat is replicating the study across different sites with different individuals and in different times. Thus, in order for the results of a survey to be externally valid, it should be generalized across settings or from one set of environmental conditions to another. This concept is also referred to in the literature as *ecological validity*.

Temporal Characteristics

This threat to external validity refers to whether the same survey results would have been obtained if the intervention were implemented (e.g., in an experimental study) and/or the data were gathered at a different point in time. In order for a survey to be characterized as externally valid, the results should be generalizable and essentially remain invariant across different points in time. Failure to consider the time variable, including possible seasonal fluctuations, can threaten the external validity of survey research.

However, there are many surveys in which the "window of time" to which their results are meant to generalize were quite brief. Examples of these surveys are those conducted for news purposes by the media on topics for which opinions well may change on a daily or weekly basis. In these cases, the temporal aspect of external validity can be an ephemeral one.

Survey Research Study Awareness

This threat to external validity refers to the possible impact on respondents of knowing that they are participating in a survey research study. This impact is known in the social science literature as the *Hawthorne effect* or *reactivity*. Thus, the participants' awareness of their participation in a survey and their thoughts about the study's purpose can influence the study's outcomes and findings. Performance, achievement, attitude, and behavior are examples of such outcomes that may be affected. The research findings may be different if the participants were unaware of their participation in the study, although this generally is not practical when using survey research. Nevertheless, the prudent researcher and researcher consumer keeps this threat to external validity in mind when deciding how well the survey results generalize beyond the study itself.

One of the methods that can be used for avoiding the awareness threat in an experimental study is by giving the participants in the control group a placebo treatment and giving the participants in the experimental group the new treatment. However, it is important to stress the fact that the researcher who uses a placebo treatment in a survey study must consider the following ethical issues:

1. Should the participants in the study be informed that some of them will be given the placebo treatment and the others will be given the experimental treatment?
2. Should the participants be informed that they will not know which treatment (placebo or experimental) they have received until the conclusion of the study?
3. Should the participants be informed that if they initially received the placebo treatment, they will be given the opportunity to receive the more effective treatment at the conclusion of the study if the results of the experimental study indicate the effectiveness of the experimental treatment?

Multiple-Treatment Interferences

This threat to external validity refers to survey research situations in which the participants are administered more than one intervention, treatment, training, testing, or surveying either simultaneously or sequentially within the same larger study. In these situations, it will be difficult to determine which one of these interventions, treatments, training, testing, and surveying is responsible for the results of the survey study unless the various treatments all are controlled with an experimental design, which often is impractical. For example, if a study were conducted within a survey to try to raise response rates and several interventions were combined (e.g., higher incentives, new recruitment scripts, special interviewer training) to form a "treatment package" for the experimental group of respondents, the researchers would not have confidence in knowing which of the interventions or their interactions brought about any observed increase in response rates. Thus, any observed effects could not generalize to other studies that did not use *all* of the same interventions in the same combination.

Pretest-Treatment Interactions

This threat to external validity refers to the effects that pretesting or pre-surveying the respondents might have on the posttest and the data gathered in the survey at the conclusion of the intervention, treatment, or program. In many types of research (e.g., longitudinal research), individuals are surveyed during a pretest to provide a baseline measure of an outcome against which the effects of the intervention can be compared at the conclusion of the study to assess the effectiveness of the intervention.

Pretesting is considered a threat to external validity because exposure to the pretest and/or pretest surveying may affect the data respondents subsequently provide in future waves of surveying. Thus, the results would be generalizable only to the population if and only if the individuals are pretested. This threat could be minimized in survey studies by extending the time lapse between the pretest and posttest administrations to decrease the pretest or pre-survey effects on the posttest or post-survey results or by using a Solomon four-group design.

Survey Research Attrition Effects

This threat to external validity refers to the effects that the dropout of the sampled individuals (i.e., attrition) might have on the results of the survey panel study and the generalizability of these results. The individuals who drop out from a multi-wave research study might have specific characteristics that are different from the individuals who did not drop out from the study, as occurs when differential attrition occurs. Consequently, the survey results would be different if all the individuals stayed in the survey study and completed the survey instruments.

The dropout threat to external validity can be avoided and minimized by keeping attrition rates as low as possible. There are many different methods to minimize attrition; for example, providing participants special incentives to encourage them to remain in the study.

Survey Nonresponse Effects

This threat to external validity refers to the effects of the individuals' nonresponse to the survey request or to certain items on the questionnaire. The respondents who do not respond at all and those who do not complete the survey instrument might have specific characteristics that are different from the individuals who respond and complete the entire survey instrument. Thus, the survey results would be different if all the individuals responded and returned the survey instrument.

The nonresponse threat to external validity may be avoided and minimized by maximizing the response rate of the survey. There are different methods for maximizing the response rates of a survey research study, including: sending follow-up surveys, sending reminder notes to return the survey, and providing different kinds of incentives. But recent research by Robert Groves and others suggests that even with high response rates, there can be considerable nonresponse error at the item level, and thus the external validity for those results would be very low.

Sema A. Kalaian and Rafa M. Kasim

See also Attrition; Cluster Sample; Coverage Error; Differential Attrition; Ecological Fallacy; Missing Data; Nonprobability Sample; Nonresponse Error; Probability Sample; Random Sampling; Simple Random Sample; Solomon Four-Group Design; Stratified Sampling; Systematic Sampling; Target Population

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Christensen, L. B. (2004). *Experimental methodology* (9th ed.). Boston: Pearson Education/Allyn & Bacon.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Leedy, P. D., & Ormrod, J. E. (2005). *Practical research: Planning and design* (8th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

EXTREME RESPONSE STYLE

Extreme response style (ERS) is the tendency for survey respondents to answer categorical rating scales in the extreme, end-most intervals, across a wide range of item content. ERS can particularly affect surveys that use Likert and semantic differential scales. ERS is a source of survey error that distorts people's true attitudes and opinions. People with relatively higher ERS will tend to have relatively high or low scores, since they tend to mark extreme intervals, while those with low ERS will tend to have more moderate scores. Thus, apparent differences in survey data and observed scores between people or groups can be an artifact caused by differences in their ERS rather than by differences in their true attitudes and opinions. ERS can also distort the relationship between variables, including survey statistics such as correlations or regression slope coefficients. Distortion from ERS increases when the mean sample response is further from the scale midpoint. ERS is positively correlated with some response styles, such as yea-saying, naysaying, response range, and standard deviation, and negatively correlated with midpoint responding.

ERS is related to demographic, personality, cultural, and national variables, which makes ERS of particular concern when making comparisons across different countries or cultures. ERS tends to increase with age and decrease with education and household income, or when a person has a more collectivist versus individual orientation. People in Southern European countries tend to have higher ERS than those in Northern European ones. ERS tends to be higher for cultures that are more masculine or that place greater emphasis on differences in power and authority.

ERS depends on characteristics of survey items. ERS tends to be higher when an item is more meaningful to respondents, is worded in the first rather than third person, or written in the respondent's primary rather than secondary language. It can also vary with the scales themselves, such as the number of intervals in the scale.

Several methods have been proposed to measure individuals' ERS and then adjust their observed survey data to compensate, as a means to remove the measurement error induced by ERS. These methods share the common goal of measuring ERS across items probing a range of uncorrelated constructs, to ensure that people's true scores on a particular construct do not unduly affect their ERS scores. One method uses a dedicated battery of items specifically designed and pretested to measure ERS. Other methods allow researchers to use item sets designed for more general survey purposes, provided the items involve several constructs. Several statistical methods have been proposed to isolate observed score variation due to ERS from variation due to differences in attitudes and opinions and other sources of response variance. These methods include structural equation modeling combined with multi-group factor analysis, item response theory, and hierarchical Bayesian ordinal regression.

Eric A. Greenleaf

See also Attitude Measurement; Item Response Theory; Likert Scale; Measurement Error; Nonsampling Error; Questionnaire-Related Error; Respondent-Related Error; Response Bias; Semantic Differential Technique; Systematic Error

Further Readings

Baumgartner, H., & Steenkamp, J. B. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143–156.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187–212.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*, 328–351.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles—Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264–277.

F

Face-to-Face Interviewing

The face-to-face interview, also called an in-person interview, is probably the most popular and oldest form of survey data collection. It has continued to be the best form of data collection when one wants to minimize nonresponse and maximize the quality of the data collected. Face-to-face interviews are often used to solicit information in projects that can be considered to be very sensitive, for example, data collection on sexual behaviors. This entry describes the advantages and disadvantages of face-to-face interviewing along with basic operational considerations for successful interviews.

Advantages

By far, the main advantage of the face-to-face interview is the presence of the interviewer, which makes it easier for the respondent to either clarify answers or ask for clarification for some of the items on the questionnaire. Sometimes, interviewers can use visual aids (e.g., so-called show cards) to assist respondents in making a decision or choice. Properly trained interviewers are always necessary lest there be problems such as interviewer bias, which can have disastrous effects on the survey data. Relatively high response rates and an almost absence of item nonresponse are also added bonuses. The opportunity for probing exists where the interviewer can get more detailed information about a particular response.

Disadvantages

In face-to-face interviewing, respondents often are not afforded the time to provide answers that might be thought out thoroughly as compared to a mail or Internet survey. Respondents essentially are expected to give an answer on the spot, which may not be very convenient or very accurate. Privacy issues continue to be a main concern in face-to-face interviews. Respondents need to be able to develop trust in the interviewer in order to disclose sensitive information. Furthermore, it is not nearly as feasible to monitor face-to-face interviews compared to what can be done with telephone interviews. The time to complete a survey project that uses face-to-face interviewing is appreciably longer than that of other data collection modes. In addition, the cost of carrying out face-to-face interviewing can be exorbitant depending on the sample size and the amount of information being collected.

Basic Operational Considerations

Advance Letter

An advance letter should be sent ahead of time so that the respondent knows when the interviewer will arrive and has sufficient time, if necessary, to prepare for the interview. If the location of the interview is somewhere other than the respondent's home, this also should be communicated clearly. The letter also must describe the study's purpose, explain the procedures to be followed, and strive to motivate cooperation. One must create a letter that is precise. A successful letter should be on one page and printed on professional letterhead. It should have a contact address, toll-free telephone number, or both, and should grasp the reader's attention in its first sentence. The aim of the letter should be cogent to the respondent. It should identify why the survey is important, why it is important to participate, who is being asked to participate, and if it is confidential or not. It should be simple to read.

Administrative and Structural Considerations

Establishing proper administration of face-to-face interviewing is integral to the success of the overall process. In particular, interviewer training and ample supervision can lead not only to higher response rates but also to the collection of higher-quality data. For example, supervisors can randomly check (validate) some interviews to ensure the reliability of the data being collected.

Types of Face-to-Face Interviewing

There are three different formats that face-to-face interviewing can take:

- Structured interviews
- Semi-structured interviews
- Unstructured interviews

Each format has particular advantages and disadvantages and is suited for particular purposes. The purpose of the study, the length of the interview, and the cost constraints are all factors to consider.

In structured face-to-face interviews the interviewer asks each respondent the same questions in the same way. This is the most basic and most common face-to-face survey type. A structured interview may include open-ended and closed-ended questions. This type of interview is usually used for large projects for which the researcher wants the same data to be collected from each respondent.

Semi-structured face-to-face interviews mainly consist of open-ended questions based on topics the researcher wants covered in the interview. Although the interview focuses on key topics, there is also the opportunity to discuss, in more detail, some particular areas of interest. The interviewer has the opportunity to explore answers more widely or other areas of discussion spontaneously introduced by the respondent. The face-to-face interviewer may also have a set of prompts to help respondents if they struggle to answer any of the questions. For example, this is particularly helpful if a respondent cannot make a decision because she or he is deficient in some knowledge of the topic.

In unstructured face-to-face interviews the interviewer has a limited number of pre-defined topics to discuss in an open-ended fashion. This type of interviewing is usually what is done in focus group environments where respondents are free to talk in a free-flowing fashion with the interviewer or moderator. The interviewer then asks questions around these topics and bases later questions on the respondent's or group's responses.

Before the Interviewing Begins

Before interviewers are sent to conduct face-to-face interviews, there are some tasks that must be accomplished. One of these is to craft an introduction to inform respondents about the interview. They may want to know about the reason(s) for conducting the survey, how the information will be used, and how the results may impact their lives in the future. The respondent usually will want to know the length of time that the interview will take. Last but not least, the respondent must be given information to trust the interviewer and the introduction should set this into motion.

Immediately before the interview, the interviewer does a check to ensure that she or he has all the equipment and materials needed. She or he should review the questionnaire, as needed. The place chosen to do the interview should be as serene as possible and be exempt from unnecessary disruptions. If respondents believe that their answers are confidential, then they will be less hesitant to respond. To increase the sense of confidentiality, names and addresses should not be placed on the questionnaires (where the respondent could see them). Instead, the researchers should use code numbers on the questionnaires and keep the names and addresses in a separate document. Also, there should be no other people present—only the interviewer and the respondent.

Incentives can improve response rates, so researchers need to decide on these beforehand. This can take the form of rewarding respondents before (noncontingent incentives) or after (contingent incentives) filling out questionnaires, or both before and after. Deployment of either form of incentive can be immediate when doing face-to-face interviewing; thus, gratification is essentially instant. Cash money is the simplest and most direct reward. It has been found consistently to be more effective when prepaid (noncontingent; e.g., sent along with an advance letter) than when promised (contingent). The larger the rewards are, the higher the response rates will be. Respondents can be rewarded with other things rather than money. Rewards can vary from ballpoint pens to movie tickets. The purposes are to express appreciation for the respondents' efforts and thus encourage participation. Prepaid incentives, in theory, contribute to the trust that should be developed between the researcher (and interviewer) and the respondent.

The Interview

The interviewer should get very comfortable and help make the respondent comfortable, possibly by using some friendly banter to establish positive rapport with the respondent. The interviewer should begin with a brief recap of the main purpose and goals of the survey project. Explanation of the confidentiality aspects of the data collection process may be the most important part of the discussion prior to the commencement of the interview.

During the face-to-face interview, the interviewer should be aware of her or his body language, as the respondent might be very suspicious of any abnormal body language. Humans communicate a great deal of information through their body language, and only a small portion of a conversational message is attributable to the words that are spoken. A higher degree of cooperation is likely if the tone of the interview is relatively informal and the respondent feels comfortable enough to give private information. An interviewer's observance of any abnormal body language on the part of the respondent should be immediately addressed during the course of the interview. During the interview, the interviewer should speak clearly and use proper grammar, but not in an overly formal or overly educated fashion.

Probing

Face-to-face interviews allow the researcher the ability to have interviewers gather detailed information about attitudes and behavior toward a particular topic. Whenever one wants to find out more about an issue or explore an answer, interviewers should be trained how to probe to gather unbiased detailed responses. Face-to-face interviewing is more compatible with gathering long and detailed answers and clarifying ambiguous answers than is telephone interviewing.

Wrapping Up the Interview

Upon finishing a questionnaire and immediately thanking the respondent, a face-to-face interviewer often will recap the information given by the respondent and reassure her or him that the information will be held in confidence. The interviewer also may remind the respondent of the use that the data will be eventually put to. The respondent also should be given the interviewer's contact details just in case there are any questions after the interviewer has left. After departing from a face-to-face interview, the interviewers should write up any relevant notes and complete any additional paperwork required by the researcher linked to the interview as soon after the interview as possible, to ensure that the interviewer can remember all the details.

Isaac Dialsingh

See also Advance Letter; Closed-Ended Question; Confidentiality; Field Survey; Focus Group; Incentives; Interviewer Monitoring; Interviewer-Related Error; Missing Data; Open-Ended Question; Probing; Respondent–Interviewer Rapport; Show Card; Validation

Further Readings

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Kahn, R., & Cannell, C. (1958). *Dynamics of interviewing*. New York: Wiley.

Factorial Design

Factorial designs are a form of true experiment, where multiple factors (the researcher-controlled independent variables) are manipulated or allowed to vary, and they provide researchers two main advantages. First, they allow researchers to examine the main effects of two or more individual independent variables simultaneously. Second, they allow researchers to detect interactions among variables. An interaction is when the effects of one variable vary according to the levels of another variable. Such interactions can only be detected when the variables are examined in combination.

When using a factorial design, the independent variable is referred to as a *factor* and the different values of a factor are referred to as *levels*. For example, a researcher might examine the effect of the factor, medication dosage, of different levels (Factor 1 with three levels: low, medium, or high) for two different types of psychotherapy (Factor 2 with two levels: Type 1 and Type 2). Because this is a form of true experiment, it requires that subjects or respondents be randomly assigned to each of the conditions.

In the literature, factorial designs are reported according to the number of variables and the number of levels in the variables. The example described in the previous paragraph is a 3×2 factorial design, which indicates that there are two factors, where Factor 1 has three levels and Factor 2 has two levels. The total number of groups (or cells or conditions) in the design is the product of the number of levels. For a 3×2 design this is six groups. In general, an $m \times n$ design has mn groups, so a 5×6 design requires 30 groups.

To make the explanation more concrete, let us consider, in detail, the simplest type of factorial design: a 2×2 design with equal numbers of people randomly assigned to each of the four groups. Suppose the researcher is testing the effect of two different forms of psychotherapy (Type 1 and Type 2) and medication dosage (low or medium) on level of symptom improvement (the dependent variable) measured on a scale of 1 (showing no improvement) to 20 (showing a great deal of improvement). Thus, there are four groups in this design to which subjects are randomly assigned: (1) Type 1 psychotherapy and low medication dosage; (2) Type 1 psychotherapy and medium medication dosage; (3) Type 2 psychotherapy and low medication dosage; and (4) Type 2 psychotherapy and medium medication dosage. The clearest way to examine the data for main effects is to put the group means in a table (see Table 1). The row and column marginals are used to examine for main effects of each of the independent variables. To examine the main effect of medication dosage on symptom improvement, the table is read across and the means are compared in the low dose row versus the medium dose row. These data show a main effect of dose with patients receiving the medium level showing greater symptom improvement compared with the low dose medication group (10 vs. 15). To examine the main effect of type of psychotherapy on symptom improvement, the table is read down by column, comparing the overall means for the two groups receiving Psychotherapy 1 versus the two groups receiving Psychotherapy 2 (15 vs. 10). The data show that patients receiving Psychotherapy 1 showed greater symptom improvement compared with patients receiving Psychotherapy 2.

Table 1 Means from a 2×2 factorial study on medication dosage and type of psychotherapy

	Psychotherapy		
	1	*2*	*Row means*
Low dose	15	5	10
Medium dose	15	15	15
Column means	15	10	12.5

To determine whether an interaction exists, we examine whether the size of the Factor 1 (type of psychotherapy) effect differs according to the level of Factor 2 (medication dosage level). If so, an interaction exists between the factors. In this example, the effect of dose differs according to which therapy the patient received. Thus, there is an interaction between type of psychotherapy and medication dosage. For patients receiving the low dose medication, there was significantly less improvement under Psychotherapy 2 than under Psychotherapy 1. But dosage level made no difference when given with Psychotherapy 1.

Because the effect of the individual variables differs according to the levels of the other variable, it is common practice to stress that any significant main effects must be interpreted in light of the interaction rather than on their own. In this example, while drug dosage showed a main effect with the medium dose leading to greater symptom improvement, on average, this effect held only for patients receiving Psychotherapy 2. The outcome of a study using a factorial design can also be depicted graphically. Figure 1 shows a bar chart and a line chart of the group means. For the bar chart, an interaction is apparent because the difference between the bars for low dosage is larger than the difference between the bars for medium dosage. For the line chart, an interaction is apparent because the lines are not parallel. Line charts should be used only with factorial designs when it makes sense to talk about

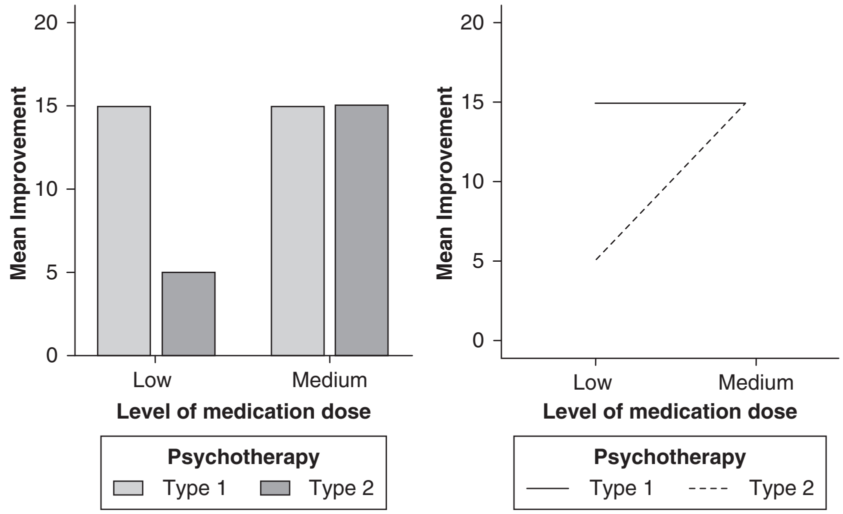

Figure 1 Graphical representations of the outcome of a factorial study: bar chart (left) and line chart (right)

intermediate values between the levels. In the present example, researchers could have given a dose halfway between low and medium; therefore, it makes sense to interpret intermediate values along the line. If the x-axis differentiated types of psychotherapy, this would probably not have been valid.

Factorial designs can take on more complex forms than that of the example presented here. They can have several factors, each with a different number of levels. Thus, a $2 \times 2 \times 3$ factorial design has three factors, two of which have two levels and one of which has three levels. This design requires 12 different groups of randomly assigned participants. With designs that have more than two factors, several different interaction effects are possible. With a three-factor design, one can have three different two-way interactions (Factor 1 with Factor 2, Factor 1 with Factor 3, and Factor 2 with Factor 3) and a three-way interaction (Factor 1 with Factor 2 with Factor 3). With these more involved factorial design studies, the nature of interactions can be determined by comparing group means to determine if a unique condition stands out or differs from other patterns of means. As the number of factors increase, it often is difficult to provide a theory to account for these higher-order interactions. Often they can arise through measurement artifacts (e.g., a floor or a ceiling effect), so caution is urged.

As the number of factors and levels increases, the number of groups also increases. A $4 \times 3 \times 4 \times 5$ design would require 240 groups. Sometimes the number of groups is too large to be practical, and rather than do a full factorial design, where there are subjects or respondents assigned to every possible condition, a *fractional factorial design* (i.e., nested design) is used, in which some groups are vacuous (empty). The choice of which conditions should be vacuous is determined by which effects the researcher does not wish to measure, and usually these are the higher-order interactions because they are often difficult to interpret. Factorial designs can also include within-subject factors, where participants take part in all levels of the factor. Further, there can be restrictions to modeling group means. Factorial designs can be used to model other statistics; in particular, when the response variable is categorical, the odds of a particular response are usually modeled.

General statistical software (such as SPSS, SAS, SYSTAT, R) will give relevant output to inform the user about the size of the main effects and interactions and whether they are statistically significant. The analysis is usually done with some form of least squares regression, often within the ANOVA framework, which has evolved with factorial design beginning with R. A. Fisher in the 1920s. There are more detailed adjustments to account for different numbers of people in groups and complex designs.

Kamala London and Daniel B. Wright

See also Analysis of Variance (ANOVA); Dependent Variable; Independent Variable; Interaction Effect; Main Effect; Marginals; Random Assignment; SAS; Statistical Package for the Social Sciences (SPSS)

Further Readings

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters: Design, innovation, and discovery* (2nd ed.). Hoboken, NJ: Wiley.
Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Factorial Survey Method (Rossi's Method)

Rossi's factorial survey method, proposed by sociologist Peter Rossi, is a technique that uses vignettes to explore individuals' beliefs and judgments. The method begins with a particular view of human nature. In this view, humans seek to know the causes of things, and they judge (evaluate) the "goodness" or "badness" of things. The drive to understand the way the world works produces positive ideas, and the drive to judge the world produces normative ideas. These positive and normative ideas can be represented by equations, termed, respectively, the *positive-belief equation* and the *normative-judgment equation*. In the positive-belief equation, also known as a "what is" equation, the individual-observer is acting as a lay scientist, whereas in the normative-belief equation, also known as a "what ought to be" equation, the individual-observer is acting as a lay judge. Rossi's factorial survey method makes it possible to estimate these equations-inside-the-head.

For example, individuals form ideas about the causes of healthiness and marital happiness, about what generates earnings and produces social harmony, and about many other matters. And they make judgments about fair compensation for workers and chief executive officers, just prison sentences, policies on trade and immigration, and so forth.

Because people differ in their life experience, social location, and information—and also in personality and culture—they may have differing perceptions about the actual world and different ideas about the just world. Thus, the positive-belief and normative-judgment equations are linked to a further equation, which describes the determinants of components of the beliefs or judgments: This equation is called a *determinants equation*.

For example, the lay scientist's view of the causes of marital happiness may be influenced by childhood observation of parental behavior, and the lay judge's view of the just prison sentence may be influenced by religious experience.

Moreover, beliefs and judgments influence many behaviors. Thus, the positive-belief and normative-judgment equations are linked to another equation, this one describing the consequences of components of the beliefs or judgments. This is called a *consequences equation*.

For example, the decision to stop smoking or the choice of a marital partner may be influenced by the positive-belief equations about healthiness and marital happiness, respectively. And the decision to participate in a strike or to make a contribution to a lobby group may be influenced by the normative-judgment equations about societal and institutional arrangements.

These four equations—(1) the positive-belief equation, (2) the normative-judgment equation, (3) the determinants equation, and (4) the consequences equation—constitute the basic set of equations in the factorial survey method. They are known, respectively, as Type II, III, IV, and V equations. (Type I is reserved for scientific approximation of the way the world works. Thus, a Type I equation represents a collective and systematic approximation to "truth," and a Type II equation represents a solitary and less explicitly systematic approximation—a Platonic "appearance" as seen by a given individual.)

The links between the four basic equations may be represented diagrammatically:

$$\text{Determinants} \rightarrow \text{Beliefs and Judgments} \rightarrow \text{Consequences}$$

Thus, the positive-belief equation and the normative-judgment equation each may join with a determinants equation to form a multi-level system of equations. Similarly, the positive-belief equation and the normative-judgment equation each may join with a consequences equation to form another (possibly multi-level) system of equations.

Note, however, that if all observers can be described by the same Type II positive-belief or Type III normative-judgment equation, then there are no determinants or consequences to study via a Type IV or Type V equation. Accordingly, the pivotal tasks in the factorial survey method are estimation of the equations-inside-the-head and assessment of the extent to which people hold the same beliefs or reach the same judgments.

Rossi pioneered the factorial survey method and developed it with several associates. Rossi's factorial survey method provides an integrated framework for estimating the positive-belief and normative-judgment equations-inside-the-head, testing for interrespondent homogeneity, and estimating the determinants and consequences equations. As comprehensively described by Guillermina Jasso in 2006, the current factorial survey framework assembles a set of tools for estimating Type II and Type III equations and carrying out the corresponding homogeneity tests and for estimating Type IV and Type V equations—incorporating such advances as seemingly unrelated regression estimators and random parameters estimators for the Type II and Type III equations and joint multi-level estimation for the pair of equations formed by a Type IV equation and a Type II or Type III equation.

All elements of the research protocol are designed with the objective of obtaining estimates with the best possible properties of the positive-belief and normative-judgment equations and the determinants and consequences equations.

Data Collection

Each respondent is asked to assign the value of a specified outcome variable (such as healthiness, marital happiness, actual wage, just wage, or fairness of an actual wage) corresponding to a fictitious unit (e.g., a person or a family), which is described in terms of potentially relevant characteristics such as age, gender, study or eating habits, access to medical care or housing, and the like. The descriptions are termed *vignettes*. One of Rossi's key insights was that fidelity to a rich and complex reality can be achieved by generating the population of all logically possible combinations of all levels of potentially relevant characteristics and then drawing random samples to present to respondents. Accordingly, the vignettes are described in terms of many characteristics, each characteristic is represented by many possible realizations, and the characteristics are fully crossed (or, in some cases, to avoid nonsensical combinations, almost fully crossed). Three additional important features of the Rossi design are (1) in the population of vignettes, the correlations between vignette characteristics are all zero or close to zero, thus reducing or eliminating problems associated with multi-colinearity; (2) the vignettes presented to a respondent are under the control of the investigator (i.e., they are "fixed"), so that endogenous problems in the estimation of positive-belief and normative-judgment equations arise only if respondents do not rate all the vignettes presented to them; and (3) a large set of vignettes is presented to each respondent (typically 40 to 60), improving the precision of the obtained estimates. The rating task reflects the outcome variable, which may be a cardinal quantity (e.g., earnings), a subjective continuous quantity (e.g., the justice evaluation of earnings), a probability (e.g., probability of divorce), a set of unordered categories (e.g., college major), or a set of ordered categories (e.g., verbal happiness assessments).

Data Analysis

The analysis protocol begins with inspection of the pattern of ratings, which, in some substantive contexts, may be quite informative (e.g., the proportion of workers judged underpaid and overpaid), and continues with estimation of the positive-belief and normative-judgment equations. Three main approaches are (1) the classical ordinary least squares approach; (2) the generalized least squares and seemingly unrelated regressions approach, in which the respondent-specific equations may have different error variances and the errors from the respondent-specific equations may be correlated; and (3) the random parameters approach, in which the respondents constitute a random sample and some or all of the parameters of the respondent-specific equations are viewed as drawn from a probability distribution. Under all approaches, an important step involves testing for interrespondent homogeneity.

Depending on the substantive context and on characteristics of the data, the next step is to estimate the determinants equation and the consequences equation. Again depending on the context, the determinants equation may be estimated jointly with the positive-belief and normative-judgment equations.

Prospects

For many years Rossi's method was somewhat difficult to implement, given the computational resources required to generate the vignette population and to estimate respondent-specific equations. Recent advances in desktop computational power, however, render straightforward vignette generation, random samples, and data analysis, thus setting the stage for a new generation of studies using Rossi's method.

Research teams around the world are exploring several new directions, building on the accumulating experience of factorial surveys on many topics and in many languages and countries. These include (a) presenting two rating tasks to the same set of respondents; (b) presenting two rating tasks to two randomly selected subsets of a set of respondents; (c) presenting two vignette worlds to randomly selected subsets of a set of respondents; (d) computerized implementation; (e) Web-based implementation; (f) pictorial vignettes and videos; and (g) systematic linkage to parallel vignette-based approaches, such as conjoint analysis and cross-national and cross-cultural measurement strategies.

Guillermina Jasso

See also Experimental Design; Random Assignment; Vignette Question

Further Readings

Berk, R. A. (1977). *Prison reform and state elites*. Cambridge, MA: Ballinger.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods and Research*, *34*, 334–423.

Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review, 42*, 639–651.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2003). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *97*, 567–584.

Li, J.-C. A. (2007). *Rethinking the case against divorce*. Unpublished doctoral dissertation, New York University.

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honor of Paul F. Lazarsfeld* (pp. 176–186). New York: Free Press.

Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach* (pp. 15–67). Beverly Hills, CA: Sage.

Rossi, P. H., & Berk, R. A. (1985). Varieties of normative consensus. *American Sociological Review*, *50*, 333–347.

Fallback Statements

Oftentimes when interviewers first make contact with a sampled respondent, the respondent is hesitant or otherwise reluctant to agree to participate in the survey. In most surveys, researchers can anticipate the nature of the concerns that will be expressed by respondents, and, as such, the researchers can train their interviewers to use targeted persuasion to try to convince the respondents to cooperate. The verbiage that interviewers are encouraged to use to persuade respondents has been termed *fallback statements*, as the interviewers must "fall back" to them in order to be successful with the respondent.

For example, respondents sampled in RDD (random-digit dialing) surveys who express reluctance to participate often will ask the interviewer, "How did you get my number?" Knowing that this question is one that commonly is asked, the researchers can provide interviewers a suggested response that they can tailor to individual respondents. A fallback statement for this particular respondent question might be something along the following lines:

> Your number was chosen by a technique called random-digit dialing. All the three-digit telephone prefixes that ring in your area were put into a computer and the computer added four more random digits to make up the seven-digit number that we used to reach you. We use this technique because it's important that we speak with people throughout your area, regardless of whether their numbers are listed or unlisted. That's the only way we can do a survey that will fairly represent the opinions of the different residents in your area.

In addition to providing interviewers with a fallback statement to help them explain how someone's telephone number or address was chosen, other topics that interviewers commonly have to explain to respondents, and thus are ones that fallback statements should address, include (a) the purpose of the survey, (b) how the data will be used, (c) additional details about how the data will be kept confidential, (c) why only one designated respondent is chosen for each household contacted, and (d) who at the survey organization can be contacted if the respondent wants to verify the legitimacy of the survey.

If any of these concerns were expressed by the majority of respondents, then it would behoove researchers to build explicit details about it into the standard introductory spiel that interviewers provide to all respondents when they first make contact. However, not every respondent has such concerns, and even among those who do, they do not necessarily share the same concerns. Thus, the survey introduction needs to be a more generic one, with the interviewers having targeted fallback statements to deploy for those respondents who express specific concerns.

Paul J. Lavrakas

See also Tailoring

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.
Lavrakas, P. J., & Merkle, D. M. (1991). *A reversal of roles: When respondents question interviewers*. Paper presented at the fifteenth annual conference of the Midwest Association for Public Opinion Research, Chicago, IL.

FALSIFICATION

Interviewer falsification, the act by a survey interviewer of faking an interview or turning in falsified results as if they were the real thing, is a well-known, long-standing, and recurrent problem that has drawn occasional attention in the research literature since the early days of the field's development. It has traditionally been referred to as *curbstoning*, a term that captures the image of the interviewer, out on field assignment, who settles on the street curbing to fill interview forms with fabricated responses instead of knocking on doors to obtain real interviews.

In recent years, the problem has drawn renewed attention because the U.S. federal government's Office of Research Integrity (ORI) made clear in 2002 that it considers interviewer falsification in any study funded by the U.S. Public Health Service to be a form of scientific misconduct. Because that designation can invoke potentially grave consequences for researchers and their organizations, a summit conference of representatives of governmental, private, and academic survey research organizations was convened in Ann Arbor, Michigan, in April 2003 by Robert M. Groves (with ORI support) to compile "best practices" for the detection, prevention, and repair of interviewer falsification. This entry draws freely on the statement generated from those meetings, which has been endorsed by the American Association of Public Opinion Research and by the Survey Research Methods Section of the American Statistical Association. This entry defines falsification, discusses its prevalence and causes, and outlines methods of prevention and control. This entry also covers actions to be taken when falsification is detected, suggests that control methods should be covered in reports of survey methods, and considers which falsification events can be handled internally.

Falsification Defined

Interviewer falsification means the intentional departure from the designated interviewer guidelines or instructions, unreported by the interviewer, which could result in the contamination of data. *Intentional* means that the interviewer is aware that the action deviates from the guidelines and instructions; honest mistakes or procedural errors by interviewers are not considered falsification. This behavior includes both fabrication (data are simply made up) and falsification (results from a real interview are deliberately misreported). It covers (a) fabricating all or part of an interview, (b) deliberately misreporting disposition codes and falsifying process data (e.g., recording a refusal case as ineligible or reporting a fictitious contact attempt), (c) deliberately miscoding the answer to a question in order to avoid follow-up questions, (d) deliberately interviewing a nonsampled person in order to reduce effort required to complete an interview, or (e) intentionally misrepresenting the data collection process to the survey management.

Prevalence, Seriousness, and Causes

Interviewer falsification is uncommon but not really rare. Because most survey organizations have practices in place to control falsification (e.g., interviewer monitoring), its prevalence is quite low. Nevertheless, any survey organization with several years of experience doing surveys is likely to have encountered one or more incidents, especially if the organization conducts face-to-face surveys that are carried out by a dispersed field staff. The consensus among practitioners is that falsification is rare in surveys that are conducted from centralized telephone facilities, because such facilities have effective monitoring controls in place to prevent and detect falsification.

In ethical terms, falsification is always a serious matter, but the seriousness of the practical consequences varies greatly. The most serious cases seem to occur in small studies by researchers who do not have professional survey services at their disposal. For example, a university professor may hire one or two graduate students to conduct all of the interviews for a small study, only to find out later that all the cases assigned to one of the students have been fabricated. The result is irreparable damage to the integrity of the research findings. In contrast, in a large study involving thousands of cases and scores of interviewers, if one interviewer is found to have falsified some of her or his cases, the relatively small number of bad cases can usually be removed without serious damage to the accuracy of the overall result.

The causes of falsification include both the characteristics of interviewers and features of the situation in which they conduct their work. As with embezzlers and others who commit breaches of trust in organizational settings, interviewers who falsify their data may have no previous records of dishonest behavior, and they sometimes are individuals suffering from financial problems, family pressures, or health issues. However, as Leo Crespi stated in a seminal 1945 article, "The Cheater Problem in Polling," cheating is not so much a problem of interviewers' morals as it is one of interviewers' morale and the factors that affect that morale. Some organizational factors that may facilitate interviewer falsification are inadequate supervision, lack of concern about interviewer motivation, poor quality control, piece-rate pay structures, and off-site isolation of interviewers from the parent organization. The design of the survey itself may play a role, as falsification seems to be more likely when interviewers are called upon to deliver cash incentives to respondents; when compensation is closely tied to production rates; when interviews are overly long, complex, or burdensome; when sample members are difficult to find; and when a too-short field period increases the pressures to produce.

Prevention and Detection

Data integrity is both a product and a reflection of organizational integrity. Researchers can enhance measures to prevent interviewer falsification by creating an organizational environment that encourages honesty, discourages falsification, enhances morale, and promotes data quality. Managers must demonstrate their commitment to data quality in both word and deed. When hiring interviewers, reference checks are recommended, and criminal background checks may be advisable despite the hiring delays and costs involved. Newly hired interviewers should be required to sign a pledge of ethical behavior that clearly states the consequences of falsification, including the threat to research and the personal consequences for the falsifier. Interviewer workloads and production goals should be realistic. The risk of falsification may increase where large incentives are offered to respondents and when interviewer pay is based on interviewer completions. These protocols are necessary in some situations, but the level of control procedures applied should reflect the increased risk.

As Groves and colleagues have outlined in their textbook, *Survey Methodology*, there are three main ways to detect interviewer falsification: (a) observational methods, (b) recontact methods, and (c) data analytic methods.

Observational Methods

Observation means that another staff member hears and/or sees the interview take place. In centralized phone facilities, observation usually takes the form of silent audio monitoring (listening to the phone interview as it takes place), sometimes coupled with video monitoring through screen capture of the computerized interview in progress. These tools significantly enhance the power of simple, direct observation of the interviewer by the floor supervisor, and where such tools are regularly employed it is virtually impossible for telephone interviewers to falsify without being promptly detected. Only a portion of the interviewing activity is monitored, but it is important that interviewers be aware that they may be monitored and not be aware of precisely when they are being monitored.

In-person field interviews have traditionally been shielded from such direct observation, but with the advent of computer-assisted personal interviewing (CAPI), the built-in microphone of the laptop computer can be programmed to digitally record random portions of the interview. These recordings can then be checked by supervising staff for indicators of falsification.

Another form of observation is to have supervisory personnel validate some or all completed interviews as soon as possible after they are completed. Experience suggests that it is very difficult to falsify data in a way that is consistent enough to appear plausible if someone else (a supervisory staff member) is looking carefully through an entire questionnaire to validate that an interview was completed properly. This form of validation is much more practical and effective in a centralized telephone facility than it is with face-to-face interviewing.

Recontact Methods

In dispersed field interviewing, the most usual means of checking for falsification is for a supervisor to communicate with the respondent after the interview is completed, to verify that the interview actually took place and that it was properly conducted. (Some telephone surveys also include verification follow-up calls, especially if the interviewers are not working from a centralized location subject to monitoring.) Recontacts can be undertaken in person, by telephone, or through the mail. Recontact methods face two challenges: (1) It can be difficult to achieve a high rate of response to the recontact attempts, and (2) it is not always clear that falsification has occurred if a respondent to the recontact should fail to recall or verify circumstances of the interview. Face-to-face recontact methods are the most expensive but generate the highest response rates; mail recontact is the least expensive but generates the fewest useful responses. Some studies mix the modes of recontact. As with monitoring, verification is undertaken only on a sample of the completed interviews, a sample that should be chosen so as to subject all interviewers to a real possibility of validation. The verification interview should go beyond simply asking if an interview takes place; it also should verify key demographic items and solicit respondent feedback on the interview experience. If a recontact fails to produce a confirmation of an interview from one of the completions submitted by an interviewer, the supervisor will typically undertake recontacts of more of that interviewer's cases, to achieve greater certainty as to whether the interviewer actually engaged in falsification.

Data Analytic Methods

Interviews and process data can sometimes be verified by examining the data record. CAPI and CATI (computer-assisted telephone interview) programs typically record interview duration and the timing of specific questions, so that complete interviews of exceptionally short duration can be flagged as suspect. Discovering an interviewer with an exceptionally high rate of production, or an unusually low rate of refusals by respondents, might lead to closer examination of that interviewer's completed interviews or call records and could lead to a decision to recontact to verify some of the contacts or interviews. Interviews could also be flagged for verification if they have unusual patterns of response, logically inconsistent responses, responses that cause the interview to skip lengthy or burdensome question sequences, or answer patterns that seem to be invariant in relation to question content (suggesting an interviewer who repeatedly selects the same response in order to speed through the fabrication process). Joe Murphy and his colleagues at the Research Triangle Institute presented a paper in 2004 describing how certain data analytic methods were deployed to detect cheating in the National Survey on Drug Use and Health. All of these techniques work best as preventives if analysis of closed and completed cases is undertaken while a survey is still in the field.

Repair and Reporting of Falsification

If falsification is suspected, survey managers should conduct an investigation by reviewing the work of the interviewer. If there is some evidence of falsification, that interviewer should be removed from data collection activities until the issue is resolved. The organization should retrospectively review all other work submitted by the suspect interviewer to check for further, previously undetected falsification. An interviewer who is found to have falsified results should be subjected to disciplinary actions under the organization's personnel policies that are appropriate to serious misconduct—in most cases, the policy is permanent dismissal.

It is obviously unethical for a survey organization to deliver data that are known to be falsified. All data known to be falsified, and any data collected by the suspect interviewer that cannot be confirmed, should be removed and—where possible—replaced with valid data from the same or from equivalently sampled cases. However, the contaminated data records need to be preserved, in a separate file from the valid results, in order to maintain an audit trail for administrative review of the misconduct episode and for any subsequent personnel actions or required reports.

Despite the potential embarrassment involved, survey organizations are obliged to report falsification episodes as an important component of understanding the quality and accuracy of a survey. The technical documentation for a survey should include a description of how monitoring or verification was carried out and a summary of the results of the efforts to detect falsification. The response rate for validation recontact efforts should be reported, along with the number of cases found to have been falsified. Efforts at cleaning, repairing, or replacing the data with valid cases should be described, and any related personnel actions should be mentioned (without naming individuals). If researchers make reports of falsification a standard part of their survey documentation, awareness of the problem and its potential solutions will be raised in the profession.

Organizations that accept funding from the U.S. Public Health Service are required to have research integrity procedures in place that require the reporting of research misconduct to the institution's research integrity officer, who is required to notify ORI in most cases and may initiate a formal inquiry. The 2003 summit, and a follow-up summit convened by Groves at Ann Arbor in 2005, sought to define a de minimus standard that would allow minor interviewer falsification incidents to be handled locally by the survey organization, using industry best practices as approved by the research integrity officer, while more serious incidents would be subject to the full scrutiny of the ORI process. The recommended threshold is that an event would exceed the de minimus standard if, in a single study, a single interviewer or a group of colluding interviewers allegedly falsifies either more than 50 interviews or more than 2% of the cases. This recommendation represents the current industry consensus. Survey organizations that follow best practices in the prevention, control, and repair of interview falsification can expect that incidents rising above this threshold of seriousness will continue to be rare.

Thomas M. Guterbock

See also American Association for Public Opinion Research (AAPOR); American Statistical Association Section on Survey Research Methods (ASA-SRMS); Data Management; Interviewer Monitoring; Interviewer-Related Error; Recontact; Survey Ethics; Validation; Verification

Further Readings

Biemer, P. P., & Stokes, S. L. (1989). The optimal design of quality control samples to detect interviewer cheating. *Journal of Official Statistics*, *5*, 23–39.

Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly*, *9*(Winter), 431–445.

Forsman, G., & Schreiner, I. (1991). The design and analysis of reinterview: An overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 279–301). New York: Wiley.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). Principles and practices related to scientific integrity. In *Survey methodology* (pp. 345–376). Hoboken, NJ: Wiley.

Interviewer falsification in survey research: Current best practices for prevention, detection and repair of its effects [Online statement of the 2003 Ann Arbor summit meeting]. (2003, April 21). Retrieved November 4, 2007, from http://www.aapor.org/uploads/falsification.pdf

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D., & Kennet, J. (2004). *A system for detecting interviewer falsification*. Paper presented at the meeting of the American Association for Public Opinion Research, Phoenix, AZ. Retrieved November 8, 2004, from http://www.rti.org/pubs/paper-detectingintrvfalsifi.pdf

Fast Busy

A fast busy is a survey disposition that is specific to telephone surveys. It occurs when an interviewer dials a number in the sampling pool and hears a very rapid busy signal. Fast busy signals are sometimes used by telephone companies to identify nonworking telephone numbers, but they occasionally occur when heavy call volumes fill all of the local telephone circuits. Telephone numbers in the sampling pool that result in a fast busy disposition usually are considered ineligible. As a result, fast busy case dispositions are considered final dispositions and typically are not redialed by an interviewer, although in some cases they may be dialed again in case the fast busy condition is only temporary.

From a telephone interviewing standpoint, the practical difference between a fast busy signal and a normal busy signal is that the pace of the fast busy signal is noticeably faster than that of a normal busy signal. It is important to note that the disposition of fast busies is different from that of busies, and thus fast busies need to have a survey disposition code that is different from the code used for normal busies. As a result, telephone interviewers need to understand the difference between busies and fast busy signals, along with the different dispositions of cases that reach normal busies and fast busy signals. This knowledge will ensure that interviewers code the fast busy cases appropriately and will prevent interviewers from making unnecessary additional call attempts on these cases.

If cases with a fast busy disposition are redialed later in the field period of a survey, it is possible that a small proportion of these numbers may no longer have a fast busy signal on the line. This may occur when a telephone company assigns the number to a new customer or puts the number (back) into service.

Matthew Courser

See also Busies; Final Dispositions; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Favorability Ratings

A favorability rating is a statistical indicator that is produced from data that typically are gathered in political polls. These ratings indicate whether the public's overall sentiment toward a politician is favorable (positive), unfavorable (negative), or neutral. Journalists often report favorability ratings as part of their coverage of political campaigns and elections.

A favorability rating about a politician is calculated by using data gathered in so-called approval questions. These questions ask poll respondents whether they "approve or disapprove of X," where X typically is the name of a politician. The favorability rating for that person is calculated by subtracting the proportion of those interviewed who say they disapprove of the person (or her or his policies, or both) from the proportion that say they approve. That is, the disapprove (negative) percentage is subtracted from the approve (positive) percentage; if there are more who disapprove than approve, then the favorability rating will be a negative number.

For example, if 65% of the people polled said they disapproved of the job George W. Bush was doing as president, while 30% said they approved (with 5% undecided), the favorability rating for Bush would be $30 - 65 = -35$. The –35 score would indicate that there are substantially more people who disapprove of the president than approve. In contrast, if 45% of those polled said they approved of Bush, 40% said they disapproved, and 15% remained undecided, the president's favorability rating would be a $+5$ $(45 - 40)$, or very slightly positive.

Favorability ratings are best understood within some comparative context. Typically this context is either to compare the favorability rating of one politician with that of another or to compare the current favorability rating of a politician with her or his previous favorability ratings. Because favorability ratings are produced by contrasting two percentages, the absolute value of the rating indicates almost nothing about the underlying dynamics of public opinion toward the politician's job performance. For example, a favorability rating of –10 can result from many varied underlying differences in public sentiment, such as 5% of the public being positive about the politician, 15% being negative, and 85% having no opinion either way; or from 55% being negative and 45% being positive. These examples are two very different circumstances and reflect great differences in what the public as a whole believes. As such, a favorability rating by itself is difficult to interpret beyond merely knowing that proportionally more people feel one way than feel the other.

Finally, as a measurement technique in polls and surveys, favorability ratings are not limited in use only to the assessment of opinions about politicians. They can be used to assess the overall valance toward any person, place, or thing. Nor do they need to be calculated from a survey item that measures the extent of approval versus disapproval that the public holds. Instead, any survey item that is evaluative—in the sense that the closed-ended response scale ranges from "good" to "bad"—can be used to compute a favorability rating.

Paul J. Lavrakas

See also Approval Ratings; Election Polls; Horse Race Journalism

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Federal Communications Commission (FCC) Regulations

The Federal Communications Commission (FCC) regulates many of the telecommunications survey and marketing researchers use through the rules under the Telephone Consumer Protection Act (TCPA), which directed the FCC to balance the fair practices of telemarketers with consumer privacy concerns.

Although some TCPA provisions apply only to commercial and sales-related communications—for example, the Junk Fax Prevention Act (JFPA), the National Do Not Call Registry, and restrictions on call abandonment and time of day—they still impact researchers. However, the TCPA restrictions on "war dialing," artificial or prerecorded messages, and cellular phone calling apply to all callers, including survey researchers.

Junk Fax Prevention Act (JFPA)

The federal JFPA amends earlier fax regulations in the TCPA to reduce the amount of unsolicited facsimile advertisements sent to businesses and residences. The law does not apply to researchers faxing surveys, collecting data via fax, or recruiting respondents via fax. The FCC defines unsolicited fax advertisements as "any material advertising the commercial availability or quality of any property, goods, or services which is transmitted to any person without the person's prior express invitation or permission, in writing or otherwise." Of course, survey researchers that fax unsolicited advertisements seeking to sell their services are bound by the JFPA.

However, individuals and businesses may send unsolicited fax advertisements to other business or residential subscribers where an *established business relationship* is present; this relationship is defined as "a prior or existing relationship formed by a voluntary two-way communication between a person or entity and a business or residential subscriber with or without an exchange of consideration, on the basis of an inquiry, application, purchase or transaction by the business or residential subscriber regarding products or services offered by such person or entity, which relationship has not been previously terminated by either party." Alternatively, fax advertisements can be sent if the recipient gives prior express consent.

All fax advertisement senders must provide a clear and conspicuous opt-out notice on the first page of the ad, and a telephone number, fax number, and a cost-free mechanism (including a toll-free telephone number, local number for local recipients, toll-free fax number, Web site address, or email address) to opt out of faxes. These numbers and cost-free mechanism must permit consumers to make opt-out requests 24 hours a day, 7 days a week.

Although survey research is outside the scope of the JFPA, professional organizations generally recommend that all researchers adopt fax policies addressing respondent opt-out requests to promote respondent cooperation.

Do-Not-Call (DNC) Registry

The FCC TCPA rules first required that companies maintain their own internal do-not-call registries; subsequently, the TCPA was amended to create a federal DNC registry operated by the Federal Trade Commission in conjunction with the FCC. Telemarketers and sellers are required to search the registry at least once every 31 days and drop from their call lists the phone numbers of consumers who have registered.

Calls placed to registered lines are allowed with prior written consent or under established business relationship rules similar to those of the JFPA.

The DNC registry does not apply to survey research calls; however, a researcher that accesses the DNC registry, for whatever reason, becomes legally bound by it—that is, responsible for scrubbing their calling lists of registrants, just like a telemarketer.

Call Abandonment

The TCPA prohibits telemarketers from abandoning more than 3% of all telemarketing calls that are answered live by a person. A call is considered abandoned if it is not connected to a live sales representative within 2 seconds of the called person's completed greeting. Although these restrictions apply only to telemarketing calls, professional associations recommend that researchers strictly limit their call abandonment rates.

Time of Day Restrictions

The TCPA restricts the time of day for sales and fund-raising calls to between 8:00 a.m. and 9:00 p.m.

(local time for the called consumer). Although researchers are exempt from such restrictions, professional associations generally recommend abiding by these restrictions as a best practice.

War Dialing

War dialing is the practice of using automated equipment to dial telephone numbers, generally sequentially, and software to determine whether each number is associated with a fax line or voice line. The TCPA prohibits anyone from doing so. However, the restriction only applies if the purpose of the call is to determine whether the line is a facsimile or a voice line. For example, calling a number already known to be a voice line for the purpose of determining if it is a working or nonworking number could be outside the scope of the TCPA.

Artificial or Prerecorded Messages

The TCPA prohibits telemarketing calls to any residential phone using an "artificial or prerecorded voice to deliver a message without the prior express consent of the called party."

Although that does not apply to survey researchers, the TCPA requires all artificial or prerecorded messages to disclose (at the beginning of the call) the identity of the business, individual, or other entity initiating the call, and if a business is responsible for initiating the call, the name under which the entity is registered to conduct business with the State Corporation Commission (or comparable regulatory authority) must be stated.

Messages must state clearly the telephone number (other than that of the auto-dialer or prerecorded message player that placed the call) or address of such business, entity, or individual. Furthermore, the telephone number provided during the disclosure may not be a 900 number or any other number for which charges exceed local or long-distance transmission charges.

Calling Cellular Phones

In only limited circumstances is it legal to call cell phones for survey research purposes.

Under the TCPA, automatic telephone dialing systems (including auto-dialers and predictive dialers) cannot be used to call a 911 line, an emergency line of a hospital, a doctor's office, a health care facility, a poison control center, a fire department, a law enforcement agency, a paging service, a cellular telephone, or any service where the called party is charged for the call, or in such a way that two or more telephone lines of a multi-line business are engaged simultaneously. The TCPA rules allow for such calls to cellular phones only in cases of emergency or where there is express consent of the called party. The FCC has acknowledged that "persons who knowingly release their phone numbers have in effect given their invitation or permission to be called at the number which they have given, absent instructions to the contrary . . . . Hence, [callers] will not violate our rules by calling a number which was provided as one at which the called party wishes to be reached." The TCPA restrictions apply to both intrastate calls (calls made and originating from within the same state) as well as interstate calls (calls from one state to another), and the FCC can impose monetary penalties for violation of this restriction.

This means that automatic dialing systems are prohibited from dialing cell phones, and there is no good faith exception for inadvertent calls to cell phones. But based on the current scope of the law, survey research calls to cell phones are not prohibited outright. If a researcher is not using an automatic system but calls cell-phone numbers manually or has consent from the called party to call his or her cell phone, that researcher may be outside the scope of the law's restrictions.

Most random-digit dialing sample providers identify and remove numbers assigned to wireless carriers from their frames. To ensure the removal of landline phone numbers that have been ported to a wireless service, NeuStar as the North American Numbering Plan Administrator, the National Pooling Administrator, and the Local Number Portability Administrator licenses a database of these ported numbers that can be used for scrubbing.

Howard Fienberg

See also Cell Phone Sampling; Do-Not-Call (DNC) Registries; Federal Trade Commission (FTC) Regulations; Number Portability; Random-Digit Dialing (RDD); Telephone Consumer Protection Act of 1991

Further Readings

Federal Communications Commission: http://www.fcc.gov
National Do Not Call Registry: https://www.donotcall.gov

Federal Trade Commission (FTC) Regulations

The Federal Trade Commission (FTC) regulates various aspects of telemarketing and the collection, use, and dissemination of personally identifiable information (PII). Under the Telephone and Consumer Fraud and Abuse Prevention Act (TSR), the FTC regulates telephone solicitation. The Children's Online Privacy Protection Act (COPPA) delineates how Web site operators (including researchers) may collect and use PII from children under the age of 13 online. Under the Controlling the Assault of Non-Solicited Pornography and Marketing Act (CAN-SPAM), the FTC regulates commercial emails. Further, the FTC writes the rules enforcing consumer financial privacy thanks to the Gramm-Leach-Bliley Act. Finally, deceptive practices are regulated according to the law that originally established the FTC.

Telemarketing and Consumer Fraud and Abuse Prevention Act (TSR)

This federal act, also know as the Telemarketing Sales Rule or TSR, established rules in 1994 to prohibit certain deceptive telemarketing activities, and it regulates sales and fund-raising calls to consumers, as well as consumer calls in response to solicitation by mail. The TSR also prohibits activities commonly known as SUGing and FRUGing. *SUGing* is the practice of selling under the guise of research, while *FRUGing* is fund-raising under the guise of research. Selling, in any form, is differentiated from survey research, and the FTC recognizes that in the TSR. Occasionally, survey research companies will offer an incentive or gift to the respondent in appreciation of his or her cooperation. Such an incentive or gift could be a cash donation to a charity, a product sample, or a nominal monetary award. But sales or solicitation is not acceptable or permitted in legitimate and professionally conducted survey research and violates federal law.

Telemarketers have various restrictions in the TSR but perhaps the best-known provisions relate to the National Do Not Call Registry. To enforce the law, the TSR allows consumers to bring private civil lawsuits in federal district courts.

Children's Online Privacy Protection Act (COPPA)

The federal COPPA, signed into law in 2000, applies to the online collection of personal information from children under the age of 13. The primary goal is to place parents in control over what information is collected from their children online. The rules spell out what a Web site operator must include in a privacy policy, when and how to seek verifiable consent from a parent, and what responsibilities an operator has to protect children's privacy and safety online. The rules cannot be sidestepped by simply including a disclaimer, making the collection of PII optional, or surreptitiously inviting children to falsify their age.

COPPA applies to operators of commercial Web sites or online services directed to children under the age of 13 that collect personal information from children, operators of general audience sites that knowingly collect personal information from children under 13 years of age, and operators of general audience sites that have a separate children's area and that collect personal information from children under 13 years of age.

PII under COPPA includes full name; physical address; email address; Social Security number; phone number; screen name revealing an email address; persistent identifier, such as a number held in cookie, which is combined with personal information; and information tied to personal information, such as age, gender, hobbies, or preferences.

The FTC applies a sliding scale approach to the practice of collecting PII from children. It balances the level of information sought from a child and the level of information needed from the child's parent. Information that the Web site operator will keep for internal purposes requires simpler consent methods than information that might be externally shared with the public or a third party.

Although COPPA applies only to children under 13, professional research and marketing associations generally recommend that researchers seek parental consent for any respondents under the age of 18.

Controlling the Assault of Non-Solicited Pornography and Marketing Act (CAN-SPAM)

The federal CAN-SPAM Act, signed into law in 2003, established commercial email distribution requirements, penalties for violation of the law, and consumers' rights to opt out of future emailings.

The law bans false or misleading header information, meaning that emails must contain accurate "From" and "To" fields and routing information, including the originating domain name and email address. It also prohibits deceptive subject lines. Commercial email must be identified as an advertisement or solicitation and include the sender's valid physical postal address, as well as explicitly state that recipients can opt out of future emails.

The law requires not only that emails include a method for recipients to opt out but also that senders strictly honor such requests within 10 days, and for at least 30 days after sending the commercial email. In addition, CAN-SPAM prohibits the sale or transfer of the opted-out email address.

The CAN-SPAM Act applies to those that distribute commercial email messages. However, survey research emails may be covered under the false or materially misleading header provision—usually not an issue for the profession, give the ethical nature of research contacts. However, researchers using email to solicit business or sell goods or services are bound by the law. Survey researchers that are recruiting or inviting respondents to participate in a survey are not legally required to abide by the opt-out provisions or email identification provisions of the CAN-SPAM Act because recruiting and taking surveys are not commercial or sales-related activities. However, as part of best practices, researchers are encouraged by professional associations to include opt-out notices in all email distributions, regardless of whether the message is commercial or noncommercial in nature.

The Gramm-Leach-Bliley Act

The federal Gramm-Leach-Bliley Act, signed into law in 1999, includes provisions regulating the privacy and security of consumer financial information, which are overseen by the FTC and a variety of financial regulatory agencies.

The law restricts the disclosure of consumers' "nonpublic personal information" by "financial institutions" and requires explicit notices to customers about information-collection and information-sharing practices. The Gramm-Leach-Bliley Act allows for consumers to opt out of having their information shared with third parties, and all financial institutions are required to provide notice and opt-out opportunity before they may disclose information to nonaffiliated third parties (with certain caveats).

The FTC defines the term *financial institution* as any institution engaged in the business of providing financial services to customers who maintain a credit, deposit, trust, or other financial account or relationship with the institution. An institution must be "significantly engaged" in financial activities to be considered a financial institution.

Thus, for researchers to acquire customers' PII from financial institutions, the institutions must either (a) provide customers notice of such disclosure and their ability to opt out of it or (b) utilize an exception in the law. Under this exception, the financial institution is still required to provide notice to its customers about its information-sharing practices, but PII can be disseminated without the opt-out provision to third parties who provide services for the financial institution—for example, survey researchers conducting research for the financial institution. To take advantage of this exception, survey researchers would have to enter into a contractual agreement with the financial institution to keep the PII confidential.

Deceptive Practices

The FTC regulates survey researchers in a broad way—breaking promises can mean breaking the law. Violating stated privacy policy can be actionable under Section 5 of the original FTC authorization act (15 U.S.C. §§ 41–58) as an unfair or deceptive trade practice, as well as under similar laws at the state level.

Howard Fienberg

See also Council for Marketing and Opinion Research (CMOR); Do-Not-Call (DNC) Registries; Email Survey; Federal Communications Commission (FCC) Regulations; FRUGing; Informed Consent; Internet Surveys; Privacy; SUGing

Further Readings

Council for Marketing and Opinion Research: http://www.cmor.org
Federal Communications Commission: http://www.fcc.gov
Federal Trade Commission: http://www.ftc.gov
National Do Not Call Registry: https://www.donotcall.gov

Feeling Thermometer

The feeling thermometer is a common survey tool used by researchers to determine and compare respondents' feelings about a given person, group, or issue. Feeling thermometers enable respondents to express their attitudes about a person, group, or issue by applying a numeric rating of their feelings toward that person, group, or issue to an imaginary scale. Using a feeling thermometer, respondents express their feelings in terms of degrees, with their attitudes corresponding to temperatures. A rating of 0, very cold, indicates that a respondent does not like a given person, group, or issue at all; a rating of 100, very warm, translates to the respondent liking that person, group, or issue very much. In general, researchers consider ratings below 50 to indicate a respondent dislikes or has a negative view of a person, group, or issue; conversely, respondent ratings above 50 are indicative of positively held feelings or attitudes. The midpoint of the feeling thermometer, 50, is reserved to indicate that a respondent's feelings toward a person, group, or issue are completely neutral: He or she does not like or dislike, approve or disapprove, have positive or negative feelings toward the person, group, or issue.

Despite the seemingly simple and straightforward concept of feeling thermometers, they are susceptible to high levels of variance due to a variety of reasons associated with how individuals respond to feeling thermometers. Studies have found that some respondents tend to be "warmer" than others in applying the scale, whereas other respondents tend to be "colder." Further, they explain that some respondents, for whatever reason, restrict their ratings to relatively small portions of the thermometer, whereas others are just more open to using the entire spectrum. Additionally, an inverse relationship has been found between respondents' levels of education and thermometer ratings, with higher ratings associated with the less educated respondents.

Feeling thermometers were first used in the 1964 American National Election Study. Because feeling thermometers were introduced in an election study, people commonly associate the use of feeling thermometers with political science research. Although political scientists do utilize feeling thermometers in a wide variety of studies, many researchers in other disciplines, including psychology and sociology, frequently employ feeling thermometers in their research as well. Beyond social sciences, feeling thermometers are often used in medical fields to allow respondents, or patients, to rate their health or health-related quality of life.

Feeling thermometers are important survey instruments because they allow researchers to gather information about the direction, as well as the intensity, of respondents' attitudes and feelings toward specific people, groups, and issues. Additionally, feeling thermometers have proven to be indispensable in longitudinal studies such as the American National Election Study because they allow researchers to observe and document how peoples' feelings and attitudes about certain public figures, groups, or issues change over time.

Shannon C. Nelson

See also Attitude Measurement; Attitudes; Attitude Strength; Opinion Question

Further Readings

Wilcox, C., Sigelman, L., & Cook, E. (1989). Some like it hot: Individual differences in responses to group feeling thermometers. *Public Opinion Quarterly*, *53*, 246–257.

Field Coding

Field coding involves the use by an in-person or telephone interviewer of a standardized listing of response options to categorize open-ended responses given by respondents to questions that provide no specific response options to the respondent. This approach differs from the administration of a closed-ended question, where the response options are read to the respondent, and differs from the administration of open-ended questions, where the response is typically recorded verbatim.

With field coding, an interviewer typically asks the respondent an open-ended question and waits for a response. As the respondent replies, the interviewer records the information into one or more of the predetermined response options. Should the respondent give an answer that is not on the interviewer's list of response options, the interviewer either must interpret the answer as close as possible to one of the predetermined response options or ask follow-up probes to clarify the response. Creating as inclusive a set of response options as possible is important, which means that the researcher must anticipate (or know from previous research studies) how the sample population might respond to the particular survey questions administered in this manner. As a last resort, when an interviewer cannot map an open-ended response to a pre-specified response option, the researcher should provide the interviewer with an "other" response choice and ask that some verbatim specification of what the respondent said be written by the interviewer.

By allowing respondents to reply in their own words, field coding techniques help to establish a dialogue between the respondent and the interviewer that more closely resembles a conversation than is typically the case with the administration of closed-ended survey questions. A positive rapport can help facilitate more sincere and detailed answers from the respondent. Additionally, the use of a predetermined set of response categories allows for greater standardization of the process than might be the case with recoding of verbatim responses, which can often be incomplete or unrelated to the actual question asked. A researcher must anticipate possible responses by the respondent, which requires development of an inclusive but mutually exclusive set of response options. Pretesting of the initial response options helps create the set of possible responses to use in the coding list.

Use of a standardized set of response options may, however, limit the capture of more complex responses, those which do not fit into the predetermined categories. In this respect, field coding may produce data that are less comprehensive than the recording of full verbatim responses or the taking of field notes by the respondent. Because the interviewer takes an active role in "creating" the respondent's response, field coding is susceptible to reactivity (i.e., changes in the respondent's answers caused by interaction with the interviewer or the setting) and to coder variance (variation in how an identical response is coded across a number of different interviewers). Further, the respondent may take a longer time to reach the same answer than he or she would have had the response options been presented as part of the question. This can increase the cost of conducting a survey.

To properly administer field coded questions, interviewers must be well trained and conscientious. For example, a respondent's answer may deviate from the wording of the response options in the questionnaire (which the interviewer sees but the respondent does not), leaving the interviewer to determine how best to code the response. The choice of a category may be selected by the interviewer based on his or her interpretation of the response, or the interviewer may ask follow-up questions to help determine the best fitting response. In doing the latter, however, the interviewer should not suggest one response over another, but rather allow the respondent to choose. The techniques and questions for probing the respondents' answers can either be standardized before the interview takes place or be determined by the interviewer. A skillful interviewer will allow respondents to express their responses within the parameters of the study.

Ryan Gibb

See also Closed-Ended Question; Codebook; Coder Variance; Coding; Field Work; Open-Ended Question; Precoded Question; Probing; Reactivity

Further Readings

Bechhofer, F., & Paterson, L. (2000). *Principles of research design in the social sciences*. New York: Routledge.

Burgess, R. G. (Ed.). (1982). *Field research: A sourcebook and field manual*. London: Allen & Irwin.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. New York: Pinter.

Field Director

In the survey research community, the title Field Director is commonly used to denote the person with overall responsibility for the data collection component of a survey that uses off-site interviewing personnel as data collectors. Not to be confused with the Principal Investigator or the Project Director, which may be positions held by other staff on the project, the Field Director role is commonly limited in functional scope to all aspects of collecting data in the field. The Field Director also may be called the Field Manager or Data Collection Task Leader. (The Field Director title is sometimes used to refer to the person in charge of data collection in a centralized telephone call center, although some consider this a less appropriate use of the term.)

A Field Director typically has overall responsibility for the preparation, staffing, training, implementation, monitoring, and controlling of the field operation and ensures the task is completed on schedule, within budget, and in accordance with the project objectives and quality specifications. On large surveys, the Field Director often is supported by in-house survey specialists, who assist with data collection preparations before collection begins, as well as Regional Supervisors and Field Supervisors. In this organizational model, the Field Supervisors oversee the field data collectors and report to Regional Supervisors, who in turn report to the Field Director. The Field Director is a senior member of the project management team and typically reports directly to the Project Director or Principal Investigator.

An effective Field Director should be a functional expert in two areas: project management and field survey methodology. Given the breadth of responsibilities, a person must be experienced and skilled in both areas.

Project management skills are needed to develop and implement the data collection plan and to monitor and control the execution of the plan. The Field Director should be proficient in developing work plans, work schedules, staffing plans, communication plans, quality plans, and budgets. He or she must have managerial skills to oversee the implementation of the various plans and the leadership skills to establish and maintain an effective and committed team of support staff. Finally, the Field Director must be skilled at monitoring and controlling the technical work, the logistics, the schedule, and the project budget. These project management skills are necessary to bring the field operation to a successful conclusion on schedule and within budget.

Survey methodology expertise is needed to ensure that the data collection plan incorporates appropriate survey operational methods, procedures, and systems that will result in a successful data collection outcome that meets project specifications and expectations within budget and time constraints. Areas of required expertise include development of instrumentation and forms, training plans, data collection protocols, validation and quality assessment procedures, and post–data collection processing of the collected data.

The Field Director must be a skilled negotiator, problem solver, manager, and team player who is able to interact effectively with other members of the project team as both a peer and a supervisor. He or she must coordinate field data collection activity with other project task managers, such as those in charge of sample design, systems and programming, weighting and estimation, and analysis. As noted earlier, the Field Director will often be responsible for a task team consisting of survey specialists, Regional Supervisors, Field Supervisors, and data collectors. Depending on the size, duration, and complexity of the data collection operation, the Field Director may assign task leaders to specific field data collection components, such as development of field manuals and training programs, instrumentation, field supply logistics, field staff recruiting, validation and quality, and so on. Overall responsibility for the completion and coordination of all tasks on schedule and within budget, however, lies with the Field Director.

Randall Keesling

See also Research Management

Field Period

The field period of a survey is the time frame during which the survey instrument is "in the field," as opposed to the time when the survey instrument is under development or review in the office. It is the period during which interviews are conducted and data are collected for a particular survey. Originally, it referred to the period of time when personal face-to-face interviews are being conducted by "field interviewers." Over the course of years, the field period has come to be regarded as the period of days or months over which data for a survey were gathered from respondents, regardless of the mode of data collection that was used.

The purpose of the survey is directly related to the field period that is established. A field period might be as short as a few hours for an overnight public opinion poll or a few days for time- or event-sensitive surveys. For surveys in which the subject is less time- or event-sensitive, the field period might extend for several weeks or months. In establishing the field period for a survey, the purpose of the survey is perhaps the most significant factor. To the extent that a survey is designed to gauge public opinion in response to a specific event or activity, a short field period is appropriate. This is often the case in political polling such as voter preference. It might also be appropriate in those surveys designed to gauge health risks and behavior, such as the impact of flu shots on influenza.

In addition to the purpose of the survey, another consideration is what events are taking place during the proposed field period. For example, surveys of high school students may be more difficult during certain times of the year, such as summer vacation and holiday breaks. Buying patterns may be quite different during the weeks immediately preceding the beginning of a school year than they are at the beginning of summer vacation. Holidays and traditional vacation periods may make contacting potential respondents more difficult and result in additional costs to reach the target population.

Administering surveys with a short field period can be more costly, in part because of the number of attempts that must be made for each completion. For computer-aided telephone interviews, this cost is related to the number of calls that must be made to reach the numbers of completions required. A short field period might also require the use of a large number of interviewers to ensure that sufficient call attempts are made. For mail surveys, there may be additional costs associated with either pre-notification or follow-up mailing to encourage completion of the survey. Shorter field periods may also increase the cost of data processing (such as coding and data transformation) needed to meet the turnaround time required for reports or public release of the results.

Because shorter field period surveys make it difficult to make multiple attempts to those households that are not contacted on the first attempt, there is a greater potential for sample bias. Statistical weighting can be used to overcome some of this bias. Longer field periods can allow more attempts to be made to those difficult-to-reach households and reduce the potential nonresponse bias that may result.

For those survey designs that require the use of interviewers, longer field periods can create the potential problem of interviewer turnover. This will require that a sufficient number of trained interviewers are readily available to ensure that the survey continues to be administered in a consistent manner throughout the field period. To address this concern, it is sometimes possible to cross-train interviewers on a number of surveys so that they can be reassigned as needed. The alternative is ongoing training and, if necessary, hiring of additional interviewers.

Every survey design must include determination of the field period during which useful data can and will be collected. Careful consideration of the possible sources of bias, additional cost, and implementation issues related to administration of the survey instrument can help ensure that the data collected will accurately reflect the opinions and concerns of the population being interviewed.

Dennis Lambries

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Field Work; Interviewer Effects; Nonresponse Bias; Response Bias; Survey Costs

Further Readings

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Eds.). (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Field Survey

The term *field* is used in survey research to refer to the geographical setting where data collection takes place. Typically this refers to in-person interviewing and thus the name, field survey.

One of the key decisions when designing a survey is the choice of the mode of data collection. Field interviewing is one of three traditional modes of survey data collection (along with telephone and mail). In field surveys, which are also referred to as face-to-face or personal-visit surveys, an interviewer visits the respondent's home or office (or another location) and conducts the interview. This entry outlines the major advantages and disadvantages of field data collection and the variations that are found in modern survey research and concludes with a brief overview of the development of present-day field surveys.

Advantages and Disadvantages

Field surveys have several advantages over other modes of data collection. Lengthy, complex instruments are easier to administer in a face-to-face interaction in which the interviewer can clarify questions, present visual aids, probe responses, and assess respondent fatigue. In countries like the United States that do not maintain a national registry of residents, selecting neighborhood blocks and listing the housing units on selected blocks, as is often done for field surveys, provides more complete coverage of the household population than do studies based on telephone numbers or mailing addresses. Response rates are typically higher in face-to-face surveys.

On the other hand, field interview costs are very high relative to other modes of data collection, sometimes 5 to 10 times those for telephone surveys. Large field data collections involve significant time for planning and implementation and require hiring, training, and supervising a large, geographically dispersed field staff. In terms of data quality, the presence of an interviewer may cause respondents to adjust their answers to survey items in order to report socially appropriate or desirable responses.

Variations

Field surveys can be implemented in a number of ways and can be used to collect a wide range of data. It is common to record interviewer observations on characteristics of the neighborhood and housing unit. In surveys that ask for sensitive information such as drug use or sexual behavior, some questions may be self-administered; that is, respondents read and answer the questions on their own either during or after the interview. For example, the National Survey on Drug Use and Health, a large annual field survey of approximately 70,000 U.S. persons 12 years old and older, which is sponsored by the U.S. Substance Abuse and Mental Health Services Administration and conducted by RTI International, uses ACASI (audio computer-assisted self-interviewing) in which respondents listen to questions using earphones and enter their responses on a laptop computer.

Field survey protocols may include the administration of tests of physical performance (e.g., walking speed, grip strength) or cognitive ability (e.g., memory tasks, word recognition) or the recording of physical measurements (e.g., height, blood pressure). Biological specimens such as blood or saliva or environmental specimens such as soil or dust may be taken as part of the in-person visit, as is done, for example, in the National Health and Nutrition Examination Survey. In mixed-mode studies, sample members may first be asked to complete the survey using mail, Internet, or telephone modes. Only those sample members who do not respond via these modes are followed up with a more expensive field survey request. In panel studies that collect data from the same persons at multiple time points, like the Current Population Survey, field interviewing may be used in the initial interview to motivate sample members to participate and report accurately. Later rounds of interviews are then completed using less expensive telephone interviews.

Development of Present-Day Field Studies

The roots of modern field survey research can be found in part in the studies of the poor carried out by Charles Booth and colleagues in London during the late 19th and early 20th centuries.

In the decades to follow, the use of field surveys grew dramatically as there were attempts to systematically record and analyze sample survey data on a variety of phenomena, from consumer preferences to unemployment. However, in the private and academic sectors, field surveys would later be replaced by mail and telephone surveys that were cheaper and expected to yield similar data based on methodological studies. Today, most national field data collections are sponsored by the federal government.

Historically, field interviews were completed using paper-and-pencil questionnaires, but by the end of the 20th century, most large field studies in the United States had transitioned to computer-assisted personal interviewing (CAPI) instruments that were administered using laptop computers. The first national household survey to use CAPI in the United States was the 1987 Nationwide Food Consumption Survey, conducted by National Analysts.

Ashley Bowers

See also Area Probability Sample; Audio Computer-Assisted Self-Interviewing (ACASI); Computer-Assisted Personal Interviewing (CAPI); Current Population Survey (CPS); Face-to-Face Interviewing; Field Work; National Health and Nutrition Examination Survey (NHANES)

Further Readings

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method* (pp. 39–78). New York: Wiley.

Groves, R. M., Fowler F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). An introduction to survey methodology. In *Survey methodology* (pp. 1–37). Hoboken, NJ: Wiley.

Groves, R. M., Fowler F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). Methods of data collection. In *Survey methodology* (pp. 137–168). Hoboken, NJ: Wiley.

Nicholls, W. L., II, & Kindel, K. K. (1993). Case management and communications for computer assisted personal interviewing. *Journal of Official Statistics*, *9*(3), 623–639.

Oishi, S. M. (2003). *The survey kit: Vol. 5. How to conduct in-person interviews for surveys* (2nd ed.). Thousand Oaks, CA: Sage.

Rothschild, B. B., & Wilson, L. B. (1988). Nationwide Food Consumption Survey 1987: A landmark personal interview survey using laptop computers. *Proceedings of the Bureau of the Census fourth annual Research Conference* (pp. 341–356).

FIELD WORK

Field work encompasses the tasks that field staff, such as field interviewers or telephone interviewers, perform before or during the data collection field period of a survey. Field work refers to both telephone and in-person studies. For telephone interviewing, field work is usually restricted to the field period. For in-person studies, field work may take place before, during, and after the field period for the study.

The field work for telephone studies usually involves working at a telephone survey center with computer-assisted telephone interviewing (CATI) stations. In RDD (random-digit dialing) surveys, the computer at the CATI station dials randomly generated telephone numbers that are programmed into the CATI system. Given that the telephone numbers are randomly generated, part of the telephone interviewer's field work is to screen the telephone numbers to determine whether they are eligible for the study sample. For instance, if a study is sampling members of households, the interviewer has to determine whether the phone number dialed reaches a household, a business, or is a nonworking number. Usually, the interviewer samples eligible respondent(s) at the household by asking the person who answers the phone questions designed to randomly select a person or persons in the household. The telephone interviewer then administers the CATI questionnaire to the sampled respondent over the telephone or makes an appointment for a callback when the designated respondent will be available. In case of a refusal, the telephone interviewer uses his or her knowledge about the study and about refusal conversion skills to convince the respondent to participate. If not successful, the interviewer records the reason for the refusal in detail so that the case can be contacted again for refusal conversion.

The field work for in-person studies is more extensive than for telephone interviewing. Field work conducted in preparation for the field period for random sampling studies can include listing of dwelling unit addresses, which is performed by *listers* or *enumerators*. These field staff members work at the selected geographical areas for the study identifying eligible units and listing their addresses for sampling. For list samples, field work may involve contacting institutions to obtain lists of employees, members, or clients for sampling.

Field work for an in-person interviewer (or enumerator) during the study's field period may include locating the sampled units on a map and planning the most efficient way to travel to the area to conduct the interviews. Once in the area, the field interviewer contacts the sampled unit to request participation in the study. Field work may involve screening households or businesses to identify eligible respondents using a screening questionnaire. Once the eligible respondents are identified, the field interviewer administers the main questionnaire to the sampled respondents usually via computer-assisted in-person interviewing (CAPI) but also sometimes via a paper questionnaire. In some studies, field work also involves administering literacy assessments, collecting samples (e.g., hair, urine), and taking other health measurements of respondents, such as height and weight, in addition to administering a CAPI questionnaire. Field interviewers also ensure that respondents fill out any required study forms such as consent forms. Field interviewers may also call respondents to schedule appointments for additional interviews or assessments as part of their field work. Field interviewers may plan and implement refusal conversion strategies to convert refusals incurred by the interviewer or transferred from another interviewer.

Other tasks that are part of field work for in-person interviewers are recording the result of contacts on a computer or case folder and submitting completed work via mail and online data uploads. Field interviewers report to a field supervisor or to a home office in person or via telephone conference call on a regular basis to discuss field work in progress. Field interviewers keep track of the hours they spend doing field work and traveling to and from the field and expenses incurred while in the field and submit those on a timely basis to their supervisor. They also keep track of their supplies to ensure that they have the study materials in the field when they need them. Field work for some interviewers may involve traveling to other sampled locations in the study to help to convert refusals or screen households.

For longitudinal studies, field work usually includes locating respondents who moved since the last wave of the study: This is known as tracking and tracing. Quality control procedures to monitor the work of phone or field interviewers are also considered part of field work.

Lillian Diaz-Hoffman

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Consent Form; Designated Respondent; Field Period; Interviewer; Longitudinal Studies; Refusal Conversion; Supervisor

FINAL DISPOSITIONS

Final dispositions (or final sample dispositions) are a set of codes or categories used by survey researchers to document the ultimate outcome of contact attempts on individual cases in a survey sample. Assigned after field work on a survey has been completed, final dispositions provide survey researchers with a terminal, or ending, status of each unit or case within the sampling pool. Survey researchers use final sample dispositions for two reasons: (1) to calculate response rates and (2) to help assess whether the sample might contain nonresponse error.

One important purpose of final dispositions is to calculate survey response rates. It is common practice for survey researchers to compute the response rates at the end of a survey's field period. Response rates are a common measure of survey quality, and typically it is assumed that the higher the response rate is, the higher the quality of the survey data is. Because the final dispositions categorize the outcome of each case (or unit) in the sampling pool, final dispositions make it possible for survey researchers to calculate survey response rates.

A second important purpose of final dispositions is to assess potential nonresponse error in the sampling pool. Correct or not, a common assumption is that there is more nonresponse error in survey samples with lower response rates than in survey samples with higher response rates. Although determining the amount of nonresponse error in survey data requires more than just the survey response rate, calculating survey response rates requires final dispositions and is an important first step in understanding whether nonresponse error is present in survey data.

Types of Final Dispositions

At the end of a survey field period, survey cases generally can be classified into four groups: (1) completed interviews, (2) eligible cases that were not interviewed or who did not complete the survey questionnaire (nonrespondents), (3) cases of unknown eligibility (some of which are likely to be nonrespondents), and (4) cases that were ineligible for the interview. In order to categorize the variety of possible outcomes of survey cases into each of these four broad categories, researchers usually use a more extensive and refined system of subcategories that are assigned to each unit in the sampling pool during the field period. In an effort to permit reporting of comparable final dispositions across survey organizations and survey projects, survey-related professional organizations such as the American Association for Public Opinion Research have developed standardized definitions for final dispositions.

Completed Interviews

Final dispositions for completed interviews divide the category into two subgroups: completed (full) interviews and partial interviews. Full interviews are interviews in which the respondent has provided data for each question in the survey instrument. The definition of a partial interview tends to vary across survey organizations but commonly includes those cases for which the respondent has provided data for a majority of questions in the survey instrument, including questions that are key variables for the purpose of the study.

Eligible Cases With No Data Gathered

Final dispositions for eligible cases from which no data were gathered divide the category into refusals, breakoffs, noncontacts, and "other cases." *Refusals* are cases in which some contact has been made with the sampled unit or named respondent, and the named respondent or a responsible member of the sampled unit has declined to participate in the interview. *Breakoffs* are cases in which data collection began, but the respondent refused or was unable to complete the interview (for in-person and telephone modes) or the questionnaire (for Internet and mail modes). The definition of *noncontact* varies depending on the mode of interviewing. For in-person interviews, a noncontact results when an interviewer is unable to gain access to a building, when no one is reached at a housing unit, or when the named respondent is away or unavailable. For telephone interviews, a noncontact results when the selected respondent is never available or when only an answering machine can be reached, but the message confirms that the telephone number is that of a household unit. "Other cases" include cases in which the respondent is located and does not refuse the interview but is unavailable or unable to complete the interview because of death, illness, physical or mental limitations, language problems or barriers, or other uncommon reasons.

Cases of Unknown Eligibility

Cases of unknown eligibility include situations in which it is not clear whether an eligible household exists and also situations in which a household unit exists, but it's not clear whether an eligible respondent is present within the household unit. For in-person interviews, cases of unknown eligibility include cases that were not attempted or worked by an interviewer, cases that could not be reached or that were in an unsafe area, and cases for which a valid address could not be located. For telephone surveys, cases of unknown eligibility include telephone numbers that are always busy, numbers in which no one ever answers, answering-machine messages that do not indicate whether the number belongs to a household unit, and technical phone problems that prevent the call from ever being completed properly. For mail and Internet surveys, unknown eligibility includes all sampled addresses from which the researcher receives neither a response nor any feedback about whether the survey invitation was ever received.

Ineligible Cases

For in-person household surveys, ineligible cases consist of household units included in the sample by error, nonresidential units, vacant households, household units with no eligible respondent, and situations where quotas have been filled. In addition, for telephone household surveys, ineligible cases include fax or data lines, nonworking numbers, or nonresidential numbers.

Converting Temporary Dispositions to Final Dispositions

At the end of a survey field period, many cases will already have reached a logical final disposition. These cases include completed interviews, refusals, and ineligible numbers, among others. However, some cases will not have reached a final disposition and will still have a temporary disposition code. (Temporary disposition codes are used to record the outcomes of contact attempts when the contact has not resulted in a final disposition.) Examples of temporary disposition codes include maximum call limit met, callback, no callback by date of collection cut-off, ring-no-answer, busy, and appointments that were not kept by the interviewer or the respondent.

Temporary disposition codes must be replaced with final case dispositions before these cases can be included in the calculation of response rates. For these cases, researchers must assign final dispositions by reviewing the pattern of disposition codes and call/contact outcomes recorded for each individual case and using this information to determine the final disposition code that "best" describes the case. (Computer algorithms can be written to make most, if not all, of these decisions.) In considering the proper final disposition code to use, survey researchers must consider the best information from all contact attempts. Because the information across contact attempts might be contradictory, three factors merit special attention: (1) the case's situation on status day (usually the first day of the field period or the first day that a case was contacted); (2) the certainty of the information on case contact attempts (information across contact attempts might be uncertain and researchers in these cases most often should take the conservative approach of assuming a case is eligible or possibly eligible unless there is reliable information to suggest otherwise); and (3) the hierarchy of disposition codes (disposition codes in which there was human contact take precedence over others, and generally in these cases, the last disposition in which there was human contact will serve as the final disposition). For example, if the last contact attempt with a sampled household results in a noncontact disposition, but a previous contact resulted in a refusal, most survey researchers would consider the final disposition of this case to be a refusal.

Matthew Courser

See also Dispositions; Nonresponse Error; Paradata; Response Rates; Standard Definitions; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author. Retrieved March 10, 2008, from http://www.aapor.org/uploads/standarddefs_4.pdf

Finite Population

Most statistical theory is premised on an underlying infinite population. By contrast, survey sampling theory and practice are built on a foundation of sampling from a finite population. This basic difference has myriad ramifications, and it highlights why survey sampling is often regarded as a separate branch of statistical thinking. On a philosophical level, the theory brings statistical theory to a human, and thus necessarily finite, level.

Before describing the basic notion of finite population sampling, it is instructive to explore the analogies and differences with sampling from infinite populations. These analogies were first described in Jerzy Neyman's seminal articles in the 1930s and are discussed in basic sampling theory textbooks such as William Cochran's in the 1970s. In the general framework of finite population sampling, we consider samples of size n from a finite population of size N, that is, a population with N elements or members.

The bridge of finite to infinite population sampling is also seen in terms of a finite population correction (fpc) that applies to the variances under most sampling designs. Finite population sampling typically begins with simple random sampling (SRS), the simplest form of sampling design, which can be considered with replacement or without replacement. For SRS designs, the fpc may be expressed as $1 - n/N$, or $1 - f$, where f is the sampling fraction or the sampling rate, $f = n/N$. Clearly, the fpc does not materially affect variances when sampling from large populations, particularly when the sample is not too large itself.

Finite population corrections are applicable for estimation but may not be necessary for many inferential uses such as statistical testing (e.g., comparisons between subgroups). In many instances, it is more sensible to consider an underlying infinite population when comparing subgroup parameters. In general, an infinite population approach to sampling has been developed that is based on superpopulation models. The superpopulation approach treats the value associated with a population unit as the realization of a random variable rather than as a fixed number.

Ronaldo Iachan

See also Finite Population Correction (fpc) Factor; *n*; *N*; Simple Random Sample; Superpopulation

Further Readings

Cassel, C., Särndal, C.-E., & Wretman, J. (1977). *Foundations of inference in survey sampling*. New York: Wiley.
Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*, 558–606.
Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*, 101–116.

Finite Population Correction (fpc) Factor

The finite population correction (fpc) factor is used to adjust a variance estimate for an estimated mean or total, so that this variance only applies to the portion of the population that is not in the sample. That is, variance is estimated from the sample, but through the fpc it is used to assess the error in estimating a mean or a total, which is due to the fact that not all data from the finite population are observed.

This concept is found throughout sample survey statistics, but this entry concentrates on the simplest of design-based sample survey statistics, simple random sampling (without replacement). A sample of n observations for a data element of interest, say, pairs of shoes sold, are randomly selected from the N members of the universe, say, of all shoe stores or dwellings, respectively, in a geographic region. (This can also be done by strata in stratified random sampling. Other strategies can be more complex. Also, this concept can be applied to ratios of totals, such as price per unit.) An estimated mean or total will be found by extrapolating from the sum of the n observations in the sample, $\sum_{I=1}^{n} y_i$, to an estimate, $\hat{T}$, of the sum of these values for the universe, $T = \sum_{I=1}^{N} y_i$, where this total is estimated for the part of the population not in the sample. (If $\hat{T}$ represents an estimate of the total, T, then we can write $\hat{T} = \sum_{I=1}^{n} y_i + \sum_{I=n+1}^{N} \hat{y}_i$. This will be considered later.) Therefore, there is an error associated with making this leap, and that is the sampling error. There are nonsampling errors to consider, such as poorly constructed survey measures, data processing errors, and reporting errors, but here we concentrate on the error due to the fact that not all data were observed, only the data for members of the sample. (Note also that there is a model-based analogy to this, but the fpc is considered to be part of a design-based approach.) Nonresponse by members of the sample can be handled in more than one way, but again, here we concentrate on simple random sampling, without regard to nonsampling error, or nonresponse. This applies straightforwardly to stratified random sampling where simple random sampling is accomplished within each stratum (group). Other designs become more complicated.

Consider the estimation of a total, $\hat{T}$, as previously shown. (Means and ratios follow from there. Here, totals are discussed.) For a stratified random sample design, survey weights are used—often adjusted to calibrate for auxiliary information or in some other way—and the finite population total is estimated within each stratum by adjusting from the sample total within that stratum, to account for the data not collected. We can consider one stratum at a time and, therefore, consider simple random sampling.

To estimate the variance of an estimated total, we use the estimated variance within the sample, and accepted practice is to apply it only to the part of the population that was not in the sample, $\sum_{I=n+1}^{N} \hat{y}_i$. This may seem odd at first, but it has a certain logic, if we ignore variance due to nonsampling error to some extent. If we can estimate variance for data within a population, it must be based on the sample data, as those are all the data available. If we consider a finite population, then the variance of the estimate of a total is due to the data that are not in the sample. In other words, error in estimating a finite population data element total will be considered as being due to failure to observe all data, and instead, estimating for some of it. Thus the estimated variance is applied only to the part of the population not sampled, assuming that the variability of the data available is the same as would be found in the data not collected. Therefore, any variance estimate for a finite population total has to be adjusted downward, because the data observed are considered to have no contribution to the variance of that estimated finite population total. Thus we regard the n observations made for a given data element (say, pairs of shoes sold) to be completely known, so the variance of the estimated total will only be derived from the $N-n$ cases in the subtotal, $\sum_{I=n+1}^{N} \hat{y}_i$, shown previously, that are not known.

The fpc factor, in the case of SRS, can be written as a single adjusting factor, applied to the estimated variance of a total, and written as $\frac{N-n}{N}$. This is the ratio of unobserved members of the finite population to total population size. It represents the fraction of the finite population to which we consider variance to be relevant here. This factor can also be written as $1-f$, where $f = \frac{n}{N}$ is the sampling fraction. Therefore, for simple random sampling, $\text{fpc} = \frac{N-n}{N} = 1-f$. It is therefore the *fraction* of the finite population that is *not sampled*. (That is, f is the sampling fraction, and $1-f$ is the fraction not sampled.) Because the fpc is literally a factor in the calculation of an estimate of variance for an estimated finite population total or mean, that estimated variance is reduced to zero if $n = N$ and approaches "full value" if $n \to 0$. This leads to the following:

$$\text{If } n \to N \text{ then fpc} \to 0.$$

$$\text{If } n \to 0 \text{ then fpc} \to 1.$$

In many sample surveys of very large populations, f is very small, and the fpc may be eliminated (i.e., considered as though $\text{fpc} = 1$). However, for a complete census, fpc = 0, and the variance of the finite population total, mean, or ratio is zero.

James R. Knaub, Jr.

Official Disclaimer: This is not an endorsement by the U.S. Department of Energy or the Energy Information Administration.

See also Elements; Finite Population; *n*; *N*; Sample; Sampling Error; Sampling Fraction; Sampling Variance; Simple Random Sample; Stratified Sampling; Survey; Universe; Weighting

Further Readings

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Foreman, E. K. (1991). *Survey sampling principles*. New York: Dekker.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1993). *Sample survey methods and theory* (Vols. 1 & 2). New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer. (See especially p. 47)

Focus Group

A focus group is a qualitative research method in which a trained moderator conducts a collective interview of typically six to eight participants from similar backgrounds, similar demographic characteristics, or both. Focus groups create open lines of communication across individuals and rely on the dynamic interaction between participants to yield data that would be impossible to gather via other approaches, such as one-on-one interviewing. When done well, focus groups offer powerful insights into people's feelings and thoughts and thus a more detailed, nuanced, and richer understanding of their perspectives on ideas, products, and policies.

This entry begins by describing the historical background of focus groups. The entry then discusses issues that researchers might consider in choosing to use focus groups, including their strengths and limitations. Next, the entry describes the types of focus groups; the steps taken to prepare for focus groups; and the analysis of, and reports pertaining to, the data gathered. Finally, the entry addresses some ethical considerations in relation to focus groups.

Background

Focus groups first appeared in academic research in the 1920s. At that time, scholars such as Walter Thurstone used group interviews to develop survey instruments. During World War II, Robert Merton and Paul Lazarsfeld used them to develop propaganda and other war time materials for the U.S. government. Between the 1950s and 1980s, focus groups became increasingly prominent in marketing and yet were rarely used in academic research. Marketers began to refer to them as *group depth interviews*, in which professionals trained in probing sources of behavior could work to uncover customers' psychological motivations.

In the 1980s, focus groups became more prominent in scholarly circles. This renewed attention was due both to work in social marketing (researching public health concerns) and the emergence of scholarly articles and books on focus groups as a method in the mid-1980s and 1990s by David Morgan, Richard Krueger, and others. Today, focus groups are common for academic research, product marketing, evaluation research, and quality improvement. In these arenas, they are used to help identify problems, to assist in the planning process, to aid the implementation of ideas and programs, and to assess data and outcomes.

Considerations for Choosing Focus Groups

Focus groups thrive in marketing because they provide a useful format to learn about people's reactions to concepts. These group conversations are most appropriate when participants, as a group, are asked to respond to stimuli and then share and compare their responses with and against others in the group. Focus groups provide insight into individuals, such as how they think and feel (as well as how often and deeply they think and feel) about ideas or products, when and under what conditions their thoughts and feelings lead to behaviors, when complicated or contradictory thoughts and behaviors emerge in response to topics or products, and how diverse groups view a specific idea or product. Moreover, this method allows researchers to assess more subtle feelings people may harbor about topics, to test pilot ideas, to shed light on previously collected data, and to aid in the construction of future large scale quantitative survey studies. In all of these instances, it is important to allow data to "emerge" freely from participants and to listen for the deeper understanding of the range of ideas.

In other situations, focus groups are not an appropriate choice for researchers. Group interviews should be avoided when participants are not comfortable with each other or with the topic, when a project requires rigorous statistical data, when consensus or emotionally charged information is desired, or when confidentiality is necessary. Additionally, focus groups should not be used when the act of holding a group, and soliciting opinions and reactions on a potentially sensitive issue, implies a commitment to a group of participants that cannot be kept (i.e., those who use this method have a special obligation to be sensitive to the suggestive "force" of this method as well as the communities with whom they work).

Relative to other qualitative methods, focus groups most closely resemble open-ended interviewing and participant observation. As in open-ended interviews, focus group moderators approach groups with a protocol of questions and encourage participants to focus on an identified topic. Unlike open-ended interviews, however, focus group moderators can be flexible with how the questions are asked and should use the conversation (as opposed to the individual interview) as the unit of analysis. Like participant observation, focus groups afford the opportunity to observe interaction among individuals and require that moderators surrender some power, at least, to the group. Unlike participant observation, though, focus groups produce large amounts of data on the researcher's specific interest in a short period of time. Two criteria, then, help researchers discern if focus groups are a good methodological choice for them relative to these closely aligned approaches: Would the research project be better off with the additional individual-level data acquired from interviews (than the group-level conversation data from focus groups)? Would the research project be better off with the contextual information afforded by naturally occurring events witnessed during participant observation (than the focused, yet less naturalistic, data gathered during a focused group conversation)?

Strengths and Limitations

As with all research methods, focus groups have both strengths and limitations. Strengths of this approach include how groups provide for exploration and discovery (to learn more about ideas or people who are poorly understood), context and depth (to discover the background behind thoughts, experiences, and differences), interpretation (to uncover how things are as they are and how they got that way), and sharing and comparing across participants (to offer and sharpen ideas and perspectives through the group process). In all of these instances, researchers benefit from listening and learning from a conversation across individuals.

The limitations of focus groups are similar to those of other qualitative methods and stem from the inherent flexibility of the group interview format. Focus groups have been critiqued for not yielding generalizable findings (as they typically employ small samples—three or four focus groups—that rarely are selected using probability sampling techniques). Focus group procedures can be viewed with suspicion, as questions are not asked the same way each time with regard to ordering or phrasing, and responses are not independent (and thus the unit of analysis becomes the group). Focus group data can be nettlesome, as the results are difficult to quantify and conclusions depend on the interpretations of researchers.

Types

Many focus group experts acknowledge that these group conversations can take several forms. Perhaps the most common type is a *full group*, in which a group of 6 to 10 participants (who are recruited because they share at least one commonality of relevance to the researcher) are gathered together and led by one moderator (possibly with the aide of a facilitator who helps with procedural aspects of the focus group) for 90 to 120 minutes. Other types of groups involve at least one derivation from this approach. *Two-way focus groups* allow for one group to watch another focus group and to discuss the observed interactions and conclusions.

Dual moderator focus groups feature two moderators in which one guides the conversation and another makes sure that all desired topics are covered. *Dueling moderator focus groups*, unlike dual moderator groups, feature two moderators that encourage these two leaders to intentionally take opposite sides on the issue under discussion (and then watch the conversation that emerges as a response from the group). *Respondent moderator focus groups* invite one or more of the participants to act as the moderator on a temporary basis in order to add another layer of perspective to the conversation. *Client participant focus groups* enable one or more clients of the group to engage in the discussion, either covertly or overtly, to add their desired perspective to the discussion. In addition to these takes on the standard format, focus groups can also feature fewer participants (*mini-groups* are composed of four or five participants), *teleconference focus groups* encourage interaction over a telephone or network, and *online focus groups* rely on computers and Internet networks to facilitate a conversation between participants.

Preparation Steps

Focus group preparation involves the following steps. First, researchers must decide what kind of people should be studied, how many groups should be conducted, what type of group plan should be adopted for each group type (e.g., per group recruited on at least one variable of interest to the researcher), and how participants will be recruited or sampled. Although it is rarely used, a probability sampling design can be used to sample participants. Often recruitment is done via telephone. It is recommended that at least three to four groups per group type be conducted.

Deciding upon how large an incentive should be offered is an important decision as offering too low an incentive will increase recruitment costs (because many people will refuse), possibly to the level where it would have been cost-effective to start out with a larger incentive in the first place. In deciding about the amount of incentive, consideration should be given to travel time and travel cost for the participants to come to the focus group facility.

Second, researchers should decide on a moderator. Moderators should not be of an age, ethnic background, or gender that might inhibit group members from participating in the conversation; must be comfortable with the reality that participants will have varying levels of comfort in speaking in front of the group; and must be mindful of their nonverbal behaviors (so as not to affect the group conversation).

Third, researchers should decide upon the desired level of structure for the group and on the scope of the protocol (also called a questioning route, topic guide, or discussion guide). Generally speaking, the focus group protocol should feature 10 to 12 questions for a 90-minute group.

Fourth, basic logistical issues of recruitment and compensation of participants must be considered. Participants should be selected on a variable of interest to the researchers, and efforts must be made to ensure that these individuals possess the desired background knowledge or experience to yield valuable data for the project (while also not having so much experience that they will silence other members of the group). Researchers should create careful screeners that outline the desired characteristics of group members. Researchers can also attempt to overrecruit participants for each group and then, after the participants have arrived to the location, selectively tell potentially problematic group members that the group is overenrolled (and thank such members and send them home with any promised compensation). While it might seem wasteful to pay an individual for not participating in the group, it can be far more costly to keep that individual in the group if there is a risk that he or she will threaten the group dynamics. It also is a good policy to invite one or two more people to participate than may be needed because of no-shows.

Fifth, moderators should attend to the best practices of facilitating the session. Sessions should be held around a round (or oval or rectangular) table. The moderator should be in a position to see all participants to help control the flow and content of the conversation, and if the session is being video recorded, the recording device should be behind the moderator. Name cards (with first names only) can be placed around the table to assign the participants to specific places and to facilitate the recognition of names through the conversation and during potential transcription.

Sixth, focus group data can be obtained in a variety of ways. Full transcription is the most costly but the most accurate means of generating a record of the group conversation and lends itself to myriad ways of content analyzing it. Other options include tape-based coding (in which researchers take notes from audio- or videotapes searching for pre-established themes); note-based coding (in which researchers rely on their field notes; in such instances the same researcher and moderator should be employed to ensure consistency across the field notes); and memory-based coding (recommended only for experienced moderators who have a strong sense of what they are looking for in group conversations).

Data Analysis and Reports

Most of the focus group analysis in the field of marketing is impressionistic and strives to understand and explain the motivations behind people's attitudes, responses, and feelings. Scholarly research advances a few more systematic approaches to analyzing data. The *grid technique* encourages scholars to create a table to summarize the responses of each group per question in order to compare answers per item across groups. Basic *coding techniques* advise researchers to note all mentions of a given code (derived from the research questions or topic of interest behind the project), whether the code was mentioned by all participants, and whether the code appeared in all of the groups conducted. *Indexing* is a procedure in which all extracts of data that are important to a theme, topic, or hypothesis are marked (and then the coder assigns index codes that allow researchers to attend to both themes in the data as well as the context of such themes).

Although there are no hard and fast rules, focus group reports generally include the following types of information: (a) a cover page, (b) an executive or top line summary, (c) a table of contents, (d) purposes and procedures, (e) results and findings, (f) summary of conclusions, (g) recommendations, and (h) an index. Most reports also feature a balance of direct quotations from the participants and a summary of the discussion.

Ethical Considerations

There are several ethical considerations with focus groups. One consideration involves judging if participants are at risk. Researchers can protect participants by providing them with a statement of informed consent (e.g., clarifying that participants are over 18 years of age and aware that they are participating in a study). Another ethical risk involves attending to basic privacy issues. Researchers can protect the privacy of their participants by restricting access to information that reveals their identities, for example, protecting identifying information, referring to participants only by their first names or pseudonyms, protecting access to the transcripts and tapes of the focus groups, removing or modifying identifying information on transcripts, protecting them against the sponsor of the group, and encouraging the moderator to remind participants not to overdisclose during group discussions. Yet another risk lies in the discussion of potentially stressful topics. Researchers can protect participants against stress by emphasizing how participation is voluntary, setting boundaries for the group conversation, preparing an information sheet with experts and sources in case the discussion raises issues the participants want to pursue in greater detail, and trying to include someone on the research team who has experience with germane areas of stress.

Sharon E. Jarvis and Laura Barberena

See also Content Analysis

Further Readings

Bloor, M., Frankland, J., Thomas, M., & Robson, K. (2001). *Focus groups in social research*. Thousand Oaks, CA: Sage.
Greenbaum, T. L. (1998). *The handbook for focus group research* (2nd ed.). Thousand Oaks, CA: Sage.
Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, CA: Sage.
Morgan, D. L. (1988). *Focus groups as qualitative research*. Newbury Park, CA: Sage.

Forced Choice

Forced choice refers to a specific format for response options in survey questionnaires. In a forced choice format, respondents are not given a specific option to reflect a "nonresponse" type choice, such as "no opinion," "don't know," "not sure," or "not applicable." Respondents must select a response choice that provides a specific answer to the survey item.

The elimination of item "nonresponse" choices in the forced choice format increases the number of survey records with responses that are usable for analysis. Survey designers use the forced choice format to encourage respondents to provide an actual response. The forced choice format is common in key survey questions, especially qualifier (screener) questions. For example, question items about household income and number of household members might use forced choice response formats in a survey of households below the poverty level so as to make certain that everyone provides an answer to allow the researchers to determine whether a given respondent is eligible or ineligible for the survey.

Interviewer-administered surveys sometimes use a more flexible version of the forced choice format where the item nonresponse choices are available for the interviewer to see, and thus to code, but are not explicitly read to respondents. This approach allows respondents to give these types of responses but only if they take the initiative to volunteer one of them. At the other extreme, some surveys require a valid response and terminate or discard the interview rather than accept a nonapplicable response.

As an example, response choices on a satisfaction survey might include a response scale of 1 to 7 where 1 is very dissatisfied and 7 is very satisfied. With a forced choice format, there would be no response choice to indicate a lack of opinion (though a response choice of 4 would indicate a neutral opinion). Some surveys will use a forced choice format with an even number of responses, such as a scale of 1 to 6 with no true midpoint included among the response options, rather than 1 to 7 in which 4 is the midpoint. This forces respondents to provide a response with a clear direction.

Although useful for some survey items, the forced choice format has disadvantages. The primary disadvantage is that it can contribute to measurement errors, nonresponse errors, or both. Whereas the forced choice format can discourage respondent laziness and encourage them to provide a thoughtful response, the requirement of a response can encourage respondents to answer a question in a way that does not truly reflect what they think and feel. Some respondents really may not know how they feel about an issue or may not know the information requested, and forcing a response would result in the collection of erroneous data. Also, by "forcing" a response by not providing a respondent a valid response option that indicates that she or he does not have an opinion or does not care to provide an answer to a specific question, the researcher may be increasing the chances that some respondents will be frustrated and offended and thus terminate their participation before they complete the questionnaire.

Jonathan Wivagg

See also Don't Knows (DKs); Response Alternatives

Further Readings

Bartlett, C. J. (1966). A forced-choice teacher attitude inventory. *Journal of Educational Measurement*, *3*(4), 296.

Hughes, G. D. (1969). Some confounding effects of forced-choice scales. *Journal of Marketing Research*, *6*(2), 223–226.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., et al. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*, 371–403.

Frame

A frame is used to identify elements in the population. Elements are the fundamental unit of observation in the survey. A frame may look very different depending on how the population of interest is defined and how its elements are defined. A well-defined appropriate frame is essential to the sampling process, the development of weights for use in analyses of survey data, the minimization of coverage error, and the understanding of what coverage error may exist. This entry describes the basic concept of a frame; the impact it has on sampling, weighting, and coverage; and how it is developed in relation to the survey population and the survey sample. It also discusses several commonly used frames and their specific issues.

A major goal of most surveys is to describe a specific population. For example, the U.S. government conducts two surveys specifically to estimate the rate of unemployment in the country each month: the Current Employment Statistics program (a survey of business establishments) and the Current Population Survey (CPS; a survey of people). Each month, the U.S. Census Bureau interviews a sample of people for the CPS. However, selecting that sample is difficult as there is no accurate, up-to-date list of people in the United States with contact information. Without such materials, it is difficult to draw a sample. But the U.S. Census Bureau can construct a frame of housing units in the country using various sources (the decennial census, building permits, etc.). Therefore, the U.S. Census Bureau defines the survey population as people living in housing units. This revised definition of the survey population is important because it allows for a better frame to be constructed. Of course, a disadvantage is that the homeless are not included, but this is judged to be acceptable to meet the goal of this survey.

Among the domains of research that use statistical techniques, survey research is unique in assigning so much importance to the source of sample units. Whereas most statisticians view sample units as a way to describe a process of interest, survey statisticians view sample units as a way to describe a population of interest. Other statisticians would only be interested in elements that are missing from the frame if the missing elements were informative of the process under study. For example, in sampling to study the effect of a certain drug, if the sample had no women, this would be a concern only if women reacted differently from men to the drug under study. The survey statistician may not be interested in whether women reacted differently to the drug but would want women proportionally represented in the sample frame so their role in the population could be described.

From Population to Frame

Surveys are often interpreted as applying to a general population, without any specific statements about time or relatively small subpopulations being excluded. This population, to which results are inferred, is often too simply defined for conducting a survey.

The next step for a survey researcher is to define a target population. This is often similar to the inferential population (population of inference) but excludes some elements that would be very difficult or costly to include on the frame. For example, many surveys exclude the homeless in order to use a housing unit frame, and many surveys exclude households without telephone service to use a telephone-based frame. Elements in the inferential population but missing from the target population should be easy to describe and note in the survey documentation. The target population can be thought of as the ideal survey frame.

The survey frame is an attempt to list the units in the target population. The frame may be a list of units. For example, a large company may conduct a survey of its employees, and the list may be readily available. Alternatively, the frame may be a set of procedures, materials, or both, to generate the sample. This is the case for telephone surveys that use random-digit dialing in the United States: A computer generates random 10-digit phone numbers using known working area codes (first three digits) and exchanges (next three digits). All elements of the target population should be represented on the frame, giving each element a nonzero probability of selection for inclusion in the survey sample. This probability of selection is calculated using the information on the frame and details about the sampling procedure. Because probability of selection is used in developing analysis weights, the accuracy of the survey frame (and complete documentation of it) is vital to drawing valid conclusions from data collected by the survey.

Differences between the target population and the survey frame are often referred to as coverage error. Elements that are missing from the survey frame are sources of survey undercoverage (e.g., a new employee missing from the list, telephone numbers in a new area code). Elements may also be on the survey frame more than once, leading to overcoverage (e.g., perhaps an employee with two names or a household that has two telephone numbers). Generally, undercoverage is a bigger problem for the survey researcher as identifying missed elements is difficult and costly. Overcoverage can usually be detected on the frame and fixed before sampling (e.g., removing recently terminated employees, removing business phone numbers), or detected during the interview and fixed by having good survey procedures (e.g., asking about the employee's status, asking about other phone numbers at the household).

Survey frame elements may need to be "mapped" to target population elements. For example, the CPS uses households as the frame elements, but the target population consists of people. In the CPS, when a household is selected for the survey, all the eligible residents of that household are selected for the survey, and one "reference" person answers the questions for the entire household. Other surveys treat the household as a primary sampling unit in a cluster sampling approach and subsamples only one person in each household.

Often, complex sampling procedures can reduce the burden of frame development. The CPS uses a stratified multi-stage cluster sampling approach: Within each state, the CPS samples counties, then blocks (within the sample counties), then households (within the sample blocks). In locations where address lists are incomplete, an area frame approach is used according to which lists of housing units within each block are required for the final stage of sampling. By selecting only a subset of counties, and then only a subset of blocks within those counties, the CPS only has to list housing units within the selected blocks.

Some Common Frames

One of the most common frames is a simple list of population elements. This type of frame is often used in surveys of well-defined populations, such as employees of a company or students at a school. As mentioned previously, use of lists can be combined with a complex sampling scheme, for example, by using a list of elementary schools in an area for the first level of sampling and then using lists of students within only the selected schools. Often, the successful use of a list (or lists) depends on the quality of the source of the list and having a relatively short amount of time between the list generation and survey data collection.

Beyond lists, there are two very common types of frames of households. An area frame consists of sampling defined geographic areas, then listing the housing units in those areas, then taking a sample of those housing units, and finally conducting the survey (usually with in-person interviewers). Area frames are used widely by large-scale surveys sponsored by governments because they provide good coverage by not depending on telephone service and not relying on potentially flawed lists (that may be out of date or incomplete). The major drawback to area frames is the cost associated with the in-person listing and face-to-face interviewing. Because of the high costs, multistage clustered sampling is often used, which reduces the precision of survey estimates (unless the sample size can be increased).

The second common type of household frame is telephone based. Telephone frames work well when (a) the vast majority of households in the population have telephone service and when (b) the exclusion of households that do not have telephone service is not expected to affect survey statistics. For example, one would expect fewer coverage problems when conducting a market research survey in the United States with a telephone frame compared to conducting a health survey in a developing nation. Random-digit dialing (RDD) in the United States uses a relatively simple frame: Generate a 10-digit number that is a valid telephone number (does not have a 0 or 1 in the first or fourth positions, among other criteria). So-called list-assisted RDD frames attempt to reduce the number of nonworking and business telephone numbers by first sampling area codes and exchanges that are known to have household telephone numbers (by using publicly available telephone books). This leaves only the last four digits to be randomly generated.

Frames of business establishments are often quite different from those of households. Lists are generally easily available from public directories or tax records. Many governments, including the U.S. government, maintain business registers. An important consideration for business surveys is the unit of analysis, as businesses can be thought of at different levels. At the establishment level, physically separate establishments are the basic elements of the survey. At the enterprise level, entire corporations are the basic elements of the survey, whether they consist of one establishment or thousands of establishments. A closely related decision is how to handle complex relationships in the business world (wholly owned subsidiaries, joint ventures, partnerships, etc.). Additionally, surveys of businesses often use sample designs that require measures of size on the frame. A commonly used measure of size in government business surveys is revenue reported on tax records.

Jeffrey M. Pearson

See also Area Frame; Coverage; Coverage Error; Current Population Survey (CPS); Elements; Overcoverage; Population; Population of Inference; Population of Interest; Probability of Selection; Sample; Sampling; Sampling Frame; Target Population; Undercoverage; Unit; Weighting

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

U.S. Census Bureau. (2002). *Design and methodology* (Current Population Survey Tech. Rep. No. 63RV). Washington, DC: Author.

Frequency Distribution

A frequency distribution is a tabular representation of a survey data set used to organize and summarize the data. Specifically, it is a list of either qualitative or quantitative values that a variable takes in a data set and the associated number of times each value occurs (frequencies).

The frequency distribution is the basic building block of statistical analytical methods and the first step in analyzing survey data. It helps researchers (a) organize and summarize the survey data in a tabular format, (b) interpret the data, and (c) detect outliers (extreme values) in the survey data set.

For example, the following are the scores of a group of 25 students on a final math exam: 83, 75, 95, 100, 83, 85, 85, 83, 98, 89, 84, 65, 95, 98, 80, 95, 89, 75, 65, 80, 89, 80, 75, 98, and 69. The students' math scores are not listed in any order to make sense of the data or to help provide a descriptive and summary statement about the scores. From these data, as they are displayed here, it is difficult to answer questions such as the following: How many students had math scores between 80 and 90? What percentage of students had a score of 70 or less? To answer the first question, the researcher has to count up all the scores between 80 and 90. How difficult or easy this task is depends on how many math scores the researcher has. To answer the second question, the researcher needs to know how many math scores are 70 or less in the data compared to higher math scores.

To summarize this data set, the researcher needs to put the data in some sort of logical order and tally the number of times each value occurs. This simple frequency distribution is called *raw* (or *ungrouped*) *frequency distribution*. The necessary steps in creating the raw frequency distribution are as follows:

- Identify the lowest and highest variable values in the data set.
- List in ascending order all single values in the data set from the lowest to highest (e.g., see the column labeled "Score" in Table 1).
- Tally the number of the times the variable values occurred (e.g., see the column labeled "Tallies" in Table 1).
- Count the number of tallies for each variable value (e.g., see the column labeled "Frequency" in Table 1).

Thus, the simple frequency distribution of the listing of the 25 students' math exam scores will look like Table 1. (Of note, this table does not contain any percentages, which could be added to the table and are what is called *relative frequency*.)

In some situations, this simple frequency distribution tabulation is unpractical, even impossible, or simply not needed by the researcher, for instance, when the variable under consideration has continuous values with decimal points (e.g., 88.5, 75.6, 94.4) instead of discrete values (e.g., 88, 75) or when the number of possible data points (values) is too large to construct such simple frequency distribution.

Table 1 Simple frequency distribution of math scores

Score	*Tallies*	*Frequency*	*Cumulative Frequency*
65	//	2	2
69	/	1	3
75	///	3	6
80	///	3	9
83	///	3	12
85	///	3	15
89	///	3	18
95	///	3	21
98	///	3	24
100	/	1	25
Total		25	

In such situations, a different kind of tabulation, based on the range (interval) of values instead of a set of single values, is used. The data values are grouped into different intervals and the number of data values that belong to each interval is determined. Thus, instead of listing single variable values and tallying the frequencies for each listed value, as was done in creating the raw frequency distribution in Table 1, the researcher could use ranges (intervals) of variable values and count the frequencies for each interval. This tabulation scheme is called *grouped frequency distribution*.

The steps involved in creating the grouped frequency distribution are as follows:

- Find the range of the data, which is the difference between the largest and smallest variable value. For the math scores example, the data range is $100 - 65 = 35$.
- Find the interval width. Divide the range from step 1 by the desired number of intervals. For the math scores example, if the researcher desired 5 intervals/groups, the interval width is $35/5 = 7$, that is, 7 is the number of numeric values in an interval.
- Determine the starting point of the lowest interval. For the math scores example, the smallest score is 65 and the starting point of the lowest interval should begin with 65.

Table 2 Grouped frequency distribution of math scores

Score Intervals	*Tallies*	*Frequency*	*Cumulative Frequency*
65-71	///	3	3
72-78	///	3	6
79-85	/////////	9	15
86-92	///	3	18
93-100	///////	7	25
Total		25	

- Determine the ending point of the lowest interval. This step involves adding the interval width to the lower boundary and subtracting 1 $(65 + 7 - 1 = 71)$. Thus, 71 is the value at which the lower interval should end.
- List all the needed equivalent intervals to include the largest value in the data set. For the math scores example, list 65–71, 72–78, . . . , 93–100.
- Tally the values within each interval.
- Indicate the frequencies from the tallied values.

The grouped frequency distribution of the previous listing of the 25 students' exam scores will look like that shown in Table 2.

Sema A. Kalaian

See also Outliers; Percentage Frequency Distribution; Relative Frequency; Variable

Further Readings

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah: NJ: Lawrence Erlbaum.

Witte, R. S., & Witte, J. S. (2004). *Statistics* (7th ed.). Hoboken, NJ: Wiley.

FRUGing

When a survey is not conducted to gather valid information but instead to stimulate fund-raising for a cause or organization, this practice is know as FRUGing ("fund-raising under the guise of research") and rhymes with "tugging." In a FRUGing solicitation, the answers a respondent gives to the putative survey are of lesser or no importance compared to the main goal of eliciting donations.

The effect of FRUGing on the survey enterprise is a deleterious one. On average, response rates to surveys have been declining. Reasons for this are varied but include the unethical practice of conducting false surveys for an ulterior motive; FRUGing is one such practice.

The full extent and effects of FRUGing calls, mailings, and interviewing is difficult to estimate, although empirical research on nonresponse in Canada has indicated that one quarter of citizens have received a FRUGing call or mailing. Market and survey research associations in several countries have taken the lead in public education, advocating for anti-FRUGing legislation and confronting organizations that conduct FRUGing.

FRUGing solicitations are different from legitimate surveys in that the questions are not designed to accurately understand respondents' beliefs and perceptions but rather to facilitate and lead up to a request for a donation. For that reason, FRUGing questions may be overly brief, simplistic, and often are biased in favor of the issues that are important to the organization behind the FRUGing and assumed to be important to those who are being "FRUGed" by that organization.

For example, imagine a FRUGing solicitation aimed at raising money to combat climate change that might ask the following question: *How much more could you do to fight the soon-to-be catastrophic and life-changing effects of global warming? (1) A little more, (2) A good deal more, or (3) A lot more*. This type of question wording obviously attempts to predispose the respondent to positively respond to the later solicitation for a donation to combat global warming. It also uses dramatic wording to play upon the concerns of those who are being FRUGed.

Because the sample that is "surveyed" during a FRUGing solicitation is likely to have strong opinions and to be skewed about the topic of the survey, any use of the data from the survey can result in intentionally misleading or biased findings, which the funding organization may then attempt to use to influence public opinion or public policy.

Survey researchers who find themselves in a situation where they are encouraged to attach a solicitation to a survey should take the opportunity to educate their client on the unethical aspects of this practice and the research consequences of such an action. Furthermore, in addition to being an unethical practice, FRUGing telephone calls are also illegal in the United States under the Federal Trade Commission's 2003 Telemarketing Sales Rule.

In Canada, FRUGing is known as SUGing ("soliciting under the guise of research"), leading to confusion in the United States, the United Kingdom, and continental Europe, where SUGing is defined as "selling under the guise of research."

Geoffrey R. Urland and Kevin B. Raines

See also Nonresponse; SUGing; Survey Ethics; Telemarketing

Further Readings

Kimmel, A. J. (2007). *Ethical issues in behavioral research: Basic and applied perspectives*. Malden, MA: Blackwell.

Market Research Association. (2007). *Research resources: For the consumer: Research abuse (sugging and frugging)*. Retrieved February 15, 2007, from http://www.mra-net.org/resources/abuse.cfm

National Archives and Records Administration. (2003, January 29). Part III: Federal Trade Commission, 16 CFR Part 310: Telemarketing Sales Rule; Final Rule. *Federal Register*, *68*, 4580–4679.

F-TEST

An *F*-test is any statistical hypothesis test whose test statistic assumes an *F* probability distribution. The *F*-test is frequently associated with analysis of variance (ANOVA) and is most commonly used to test the null hypothesis that the means of normally distributed groups are equal, although it can be used to test a variety of different hypotheses. The *F*-test was devised as an extension to the *t*-test: *F* is equal to the squared value of *t* ($t^2 = F$). Although the *F*-test produces the same information as the *t*-test when testing one independent variable with a nondirectional hypothesis, the *F*-test has a distinct advantage over the *t*-test because multiple independent groups can easily be compared. Survey researchers often use the *F*-test because of its flexibility to compare multiple groups and to identify whether the relationship they are studying among a set or combination of independent variables has occurred by chance.

For example, if a survey researcher hypothesizes that confidence in government varies between two groups of persons with different levels of education (e.g., those with a college degree and those without a college degree), a *t*-test and an *F*-test would produce the same results. More often, one is interested in comparing multiple or subsets of independent variables. The *F*-test gives researchers the ability to examine the independent (main) effects of education and the combined (main) effects of a set of socioeconomic status (SES) variables (e.g., education, income, and occupation) as well as the potential effects of the interaction among these variables on confidence in government.

F-tests are also often used to test the effects of subsets of independent variables when comparing nested regression models. For instance, the researcher could compare the *F*-tests from a model with only the SES variables, a model with a set of variables measuring satisfaction with government services (e.g., police, fire, water, and recreation), and an overall model with both sets of variables to determine whether, as a group, the SES and government services variables make a statistically significant contribution to explaining differences in confidence in government.

The *F*-test compares the observed value to the critical value of *F*. If the observed value of *F* (which is derived by dividing the mean squared regression by the mean squared error) is larger than the critical value of *F* (obtained using the *F*-distribution table), then the relationship is deemed statistically significant and the null hypothesis is rejected. There are two types of degrees of freedom associated with the *F*-test: The first is derived by subtracting 1 from the number of independent variables and the second by subtracting the number of independent variables from the total number of cases. In output tables from statistical software packages, such as SPSS, SAS, or STATA, the *F* value is listed with the degrees of freedom and a *p*-value. If the *p*-value is less than the alpha value chosen (e.g., $p < .05$), then the relationship is statistically significant and the null hypothesis is rejected. It is important to note that the *F*-test is sensitive to non-normality when testing for equality of variances and thus may be unreliable if the data depart from the normal distribution.

Kelly N. Foster and Leah Melani Christian

See also Alpha, Significance Level of Test; Analysis of Variance (ANOVA); Independent Variable; Interaction Effect; Main Effect; Null Hypothesis; *p*-Value; Regression Analysis; SAS; Standard Error of the Mean; Stata; Statistical Package for the Social Sciences (SPSS); *t*-Test

Further Readings

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Rossi, P. H., Wright, J. D., & Anderson, A. B. (Eds.). (1983). *Handbook of survey research*. Orlando, FL: Academic Press.

G

Gallup, George (1901–1984)

One of the pioneers of the polling field in the United States, George Gallup became the living personification of the industry during the first half century of its development. He was a trained research methodologist who also had a flair for promotion, and he parlayed both into a series of the most successful polling firms in the United States. As an undergraduate, he was interested in journalism and became a student reporter and editor. His interest in, and understanding of, the newspaper business was instrumental in the development of his polling business.

George Gallup was born on November 18, 1901, in Jefferson, Iowa. He earned three degrees from the University of Iowa, including a Ph.D. in psychology. He had a strong interest in audience and attitude research, and his dissertation involved the development of a new technique for measuring newspaper readership. This work resulted in academic positions at Drake University and Northwestern University, but in 1932 he moved to New York to join Young & Rubicam as its research director and also to serve on the journalism faculty at Columbia University. His early work in New York and Princeton focused on market research designed to improve the quality of newspapers and magazines based upon the preferences of their readers, and by 1937 he was working full time in the advertising research business.

Even at this early stage of his career, he developed an interest in social and political issues as well as elections. This started when he was a student and was enlisted to help his mother-in-law run for statewide office in Iowa. She won her first election by the narrowest of margins, but Gallup began to survey her constituents and used the resulting information to help her build increasing electoral margins. He was known as a person of high ethical standards, and as his political polling work expanded and became more public, he stopped voting in presidential elections so his published polls would be free of any allegations of personal preference or bias.

During this period, he formed the American Institute of Public Opinion, from which he began to conduct national surveys of public opinion and produce a newspaper column. He also founded the Audience Research Institute where he did work on the response of film audiences to new releases, including the development of an innovative method to measure consumer reactions to films and new products at the Mirror of America, a converted theater in Hopewell, New Jersey. He became a vice president of Young & Rubicam and served in that capacity until 1947 when he turned full time to the other businesses he had developed. He managed three firms in Princeton, New Jersey. The American Institute of Public Opinion conducted the Gallup Poll and produced three syndicated newspaper articles a week from it. A second firm, Gallup and Robinson, conducted market research for a number of clients. In addition, a third firm, the Gallup Organization, conducted special surveys tailored to the interest and needs of individual clients.

Gallup catapulted to fame in 1936 because of his belief that he could apply face-to-face interviewing with well-designed quota samples to produce better estimates of election outcomes than could *The Literary Digest*. He was so confident of the superiority of these methods that he offered his main client, *The Washington Post*, a money-back guarantee if he did not outperform the magazine's mail survey. When he proved to be more accurate than *The Literary Digest* in correctly predicting a Roosevelt victory in 1936, his business was off and running.

Gallup had good results from his 1940 and 1944 pre-election polls, but disaster struck in 1948 when he and several other pollsters estimated that Thomas Dewey would beat Harry Truman. A number of factors were identified as contributors to the error, including quota sampling and the use of mailed questionnaires to interviewers and their return by mail, which meant stopping interviewing too early. Gallup resolved to devote additional time and resources to improving the methodology of pre-election polls. In addition to adopting probability sampling methods, the Gallup Organization also pioneered techniques for identifying likely voters; these techniques became widely used in the polling industry. The Gallup Organization never made an incorrect call of the winner in an American presidential election after 1948, and it had one of the most accurate records of estimation in the industry.

At heart, Gallup was a populist who believed in the civic function of polls and their ability to serve as plebiscites between elections. He was an advocate for measuring the "voice of the people" and making it known to elected officials and other policy makers. From the start, he wrote extensively on the role and function of public polls in a democratic society, starting with *The Pulse of Democracy: The Public Opinion Poll and How It Works*. His written work promoted the polling method as much as specific findings, and it generally served to propel the development of the field by increasing its visibility and potential.

In addition to his central role in the American polling industry, Gallup also developed a network of foreign associates under the umbrella of the Gallup International Research Institutes. They were at one time responsible for conducting polls in more than 70 countries overseas. He was involved in the founding of the Roper Center at Williams College and in the National Council of Public Polls. He received numerous honors for his work including the AAPOR Award from the American Association for Public Opinion Research and election to the Advertising Hall of Fame and the Market Research Hall of Fame, as well as several honorary degrees. He died in Tschingel, Switzerland, on July 27, 1984.

Michael Traugott

See also Election Polls; Gallup Poll

Further Readings

Gallup, G. H. (1964). *The miracle ahead*. New York: Harper & Row.

Gallup, G. H., & Rae, S. F. (1940). *The pulse of democracy: The public-opinion poll and how it works*. New York: Simon & Schuster.

Gallup Poll

The Gallup Poll is the longest continuous measure of public opinion in the United States, having been conducted for more than 70 years, and is the most widely recognized brand name in the field of survey research.

On Sunday, October 20, 1935, George Gallup officially launched his "scientific" polling operation nationwide with "America Speaks: The National Weekly Poll of Public Opinion." About three dozen newspapers carried his first release, including *The Washington Post*, whose editor heralded the event by hiring a blimp to pull a streamer over the city to announce the new column. Gallup called his operation the American Institute of Public Opinion, which he located in Princeton, New Jersey, where he also lived. To attract subscribers, he made a money-back guarantee that his poll-based prediction of the 1936 presidential election would be more accurate than that of *The Literary Digest*, which had correctly predicted Herbert Hoover's win in the 1928 election within less than one percentage point of the election outcome. Gallup made good on his promise, predicting Franklin Delano Roosevelt would beat Alf Landon, while the *Digest*'s final poll predicted a Landon landslide.

Gallup kept the name American Institute of Public Opinion for more than 20 years, but within a very short time, his poll was known simply as the Gallup Poll. He, too, used that name, giving souvenir cards to cooperative respondents with the announcement, "You have been interviewed for THE GALLUP POLL—The American Institute of Public Opinion."

The Gallup Poll increased its newspaper subscribers substantially over the years, though it suffered a minor setback after the 1948 election, when Gallup and almost all of the other scientific pollsters of the day predicted Thomas Dewey to beat Harry Truman. By the 1950s, Gallup had more than 200 newspaper subscribers. In 1963, he encountered his first serious competitor, Louis Harris, who began syndication of his own column. For almost a decade and a half, the Harris Poll and the Gallup Poll were the two competing sources for news about American public opinion.

In the 1970s, the major news media organizations began forming their own public opinion polls, and by the 1980s subscriptions to the Gallup Poll had fallen considerably. Gallup died in 1984, and 4 years later, his organization was bought by a small research company in Lincoln, Nebraska, called Selection Research, Inc. (SRI). By this time, Gallup polls were rarely covered in the national news media. The president and chief executive officer of this new SRI-owned Gallup Organization negotiated with CNN to form a media partnership to cover the 1992 election campaign season. They included CNN's occasional polling partner, *USA Today*, and for the first time in Gallup's history, the poll was no longer completely independent. But the new CNN/*USA Today*/Gallup Poll gained what Gallup had mostly lost after the emergence of the media polls: immediate nationwide dissemination of Gallup Poll results.

The partnership worked to all the partners' satisfaction in 1992 and was renewed several times. In 2006, the Gallup Organization refused to renew the partnership with CNN, resulting in a messy public break-up. Gallup continues to partner with *USA Today*.

In the early years of polling, George Gallup supported researchers from around the world who were interested in establishing their own polls. He freely allowed the use of his name, if researchers thought it might help gain credibility in their own countries. In 1947, he helped found the Gallup International Association, originally with 11 members and him, though today the organization has close to 60 members and interviewing capabilities in more 100 countries. That generosity posed problems for the SRI-owned Gallup as it expanded its polling business overseas. In many countries, the U.S.-based Gallup Organization could not use "Gallup Poll," because the name was owned, or claimed, by another polling organization. In several countries, the U.S.-based Gallup Organization was able to buy back its name or get court orders to allow it to legally reclaim sole ownership of the "Gallup Poll" name. But the Gallup International Association remains a viable organization. A Gallup Poll in the United States clearly refers to the original operation founded by George Gallup. But reports from "Gallup International," and even from a "Gallup Poll" in some countries, are not necessarily from the U.S.-based Gallup Organization.

David W. Moore

See also Gallup, George; Poll; Pollster; Public Opinion; Public Opinion Research

Further Readings

Converse, J. (1987). *Survey research in the United States: Roots and emergence 1890–1960*. Berkeley: University of California Press.

Gallup: http://www.gallup.com

Gallup, G. H. (1944). *A guide to public opinion polls*. Princeton, NJ: Princeton University Press.

Gallup, G. H., & Rae, S. F. (1940). *The pulse of democracy: The public-opinion poll and how it works*. New York: Simon & Schuster.

Gallup International Association: http://www.gallup-international.com

Moore, D. W. (1995). *The super pollsters: How they measure and manipulate public opinion in America* (2nd ed.). New York: Four Walls Eight Windows.

Ohmer, S. (2006). *George Gallup in Hollywood*. New York: Columbia University Press.

Rich, W. (1939). The human yardstick. *The Saturday Evening Post*, January 21.

Gatekeeper

A gatekeeper is a person who stands between the data collector and a potential respondent. Gatekeepers, by virtue of their personal or work relationship to a respondent, are able to control who has access, and when, to the respondent. Furthermore, they may be encountered on both field (in-person) and telephone data collection surveys. They may also be encountered in mail surveys in which a respondent's material must be sent to, or in care of, another individual for distribution to the respondent (e.g., sending materials to a parent for distribution to a respondent away at college or in the military, or sending materials to an employer for distribution to sampled employees).

Gatekeepers can take many forms, including guards or doormen at secured residential or business complexes; secretaries, administrative assistants, or office managers in businesses; family members, housekeepers, and so forth. For studies in which children are the intended respondents, parents may be considered gatekeepers in that their consent must be obtained for the child's participation in the study.

A single respondent may have multiple gatekeepers that must be contacted by the data collector. In a secured apartment complex, the security guard may prevent access to an individual apartment unless the data collector has obtained permission from the complex manager. The latter is now a second gatekeeper who must be convinced to grant access to the selected housing unit. If successful there, the data collector may then encounter a third gatekeeper at the housing unit in the form of a parent, other family member, or housekeeper.

It is important to consider that a single gatekeeper may control a data collector's access to not just one but many respondents. For example, in the case of an area probability sample, a security guard at a large apartment complex may prevent access to multiple sampled housing units. An establishment survey wishing to sample multiple employees at a large company may have an administrative assistant standing in the way of gaining access to the director of human resources for the company (who could also be considered a secondary gatekeeper).

Regardless of their relationship to a respondent, gatekeepers must be successfully yet carefully negotiated in order to further the research objectives. Data collectors must walk a fine line between giving gatekeepers enough information about the survey and its sponsorship to motivate them to grant access to the respondent while, at the same time, not revealing sensitive information that could violate the respondent's privacy or reflect negatively on the person. Further, data collectors must be cognizant and respectful of all local laws and regulations regarding trespassing, solicitation, and so on.

Data collectors must assess each controlled access situation and note as many relevant details as possible. For example, when refused entry by a security guard, data collectors can note the guard's name or demographics and then plan to return when another, possibly more cooperative guard is on duty. They can check locked building entrances at other times of the day when they might be open to the public. They can also walk around gated communities to determine if a pedestrian entrance may be open. Data collectors can honk a vehicle horn at a single family housing unit when an unleashed dog prevents access to the front door. If at home, the resident will likely come to the door in response.

If study protocols permit, data collectors can ask the security guard or complex manager, in lieu of granting access to the housing units, if they would distribute a study brochure or data collector name and number to the selected housing units. If asked, they may also identify which, if any, of the selected housing units are currently vacant.

As society becomes increasingly security-minded in this age of identity theft, terrorism, and crime, the presence of gatekeepers will be encountered more and more frequently. It is critical, therefore, that researchers recognize this trend, as well as the potential effect on nonresponse that gatekeepers represent. In doing so, researchers should include in their procedural manuals and interviewer training programs material on how to deal effectively with gatekeepers. Specific instructions should be included regarding what information may be shared with gatekeepers about the study and what cannot be shared in order to protect respondent confidentiality. Strategies and tools for dealing with gatekeepers should be developed, such as informational brochures suitable for gatekeepers, main office/research director contact information, letters from the research director to the gatekeeper, and so on.

Randall Keesling

See also Contactability; Controlled Access; Establishment Survey; Interviewer Productivity; Interviewer Training

General Social Survey (GSS)

The National Data Program for the Social Sciences of the National Opinion Research Center (NORC), University of Chicago, is a social indicators and data diffusion program. Its basic purposes are (1) to gather and disseminate data on American society in order to (a) monitor and explain societal trends and constants in attitudes, behaviors, and attributes, and (b) examine the structure and functioning of society in general and the role of various subgroups; (2) to compare the United States to other societies in order to place American society in comparative perspective and to develop cross-national models of human society; and (3) to make high-quality data easily and quickly available to scholars, students, and others. These goals are accomplished by the regular collection and distribution of the NORC General Social Survey (GSS) and its allied surveys in the International Social Survey Programme (ISSP).

Origins

Two social science movements in the 1960s spawned the GSS. First, the social indicators movement stressed the importance of measuring trends and of adding noneconomic measures to the large repertoire of national accounts indices. Second, scholarly egalitarianism was advocating that data be made available to scientists at all universities and not restricted to elite senior investigators at large research centers. In 1971, these ideas were presented together in a modest proposal to the National Science Foundation (NSF) for "twenty-some questions" that called for the periodic asking of items on national samples with these data immediately distributed to the social science community for analysis and teaching. Approval from NSF plus supplemental funding from the Russell Sage Foundation spawned the first GSS in 1972.

Growth

From 1972 to 2004, the GSS conducted 25 independent, cross-sectional, in-person surveys of adults living in households in the United States, and in 1982 and 1987, it carried out oversamples of African Americans. There are a total of 46,510 respondents. During most years until 1994 there were annual surveys of about 1,500 respondents. Currently about 3,000 cases are collected in a biennial GSS.

Additionally, since 1982 the GSS has expanded internationally. The cross-national research started as a bilateral collaboration between the GSS and the Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) of the Zentrum für Umfragen, Methoden, und Analysen in Germany in 1982 and 1984. In 1984, they joined with the British Social Attitudes Survey of the National Centre for Social Research and the National Social Science Survey at Australian National University to form the ISSP. Along with institutes in Italy and Austria, the founding four fielded the first ISSP in 1985. ISSP surveys have been collected annually since that time, and there are now 41 member countries (the founding four plus Austria, Belgium, Brazil, Bulgaria, Canada, Chile, the Czech Republic, Cyprus, Denmark, Dominican Republic, Finland, France, Hungary, Ireland, Israel, Italy, Japan, Korea (South), Latvia, Mexico, the Netherlands, New Zealand, Norway, the Philippines, Poland, Portugal, Russia, Slovakia, Slovenia, South Africa, Spain, Sweden, Switzerland, Taiwan, Turkey, Uruguay, and Venezuela).

Content

The GSS lives up to its title as "General." The 4,624 variables in the 1972–2004 cumulative data set run from ABANY (legal abortion if a woman wants one for any reason) to ZOMBIES (behavioral medication for children) and have core batteries on such topics as civil liberties, confidence in institutions, crime/violence, gender roles, government spending, intergroup relations, psychological well-being, religion, and work.

The balance of components has changed over time, but currently half of the GSS is replicating core topics, one sixth deals with cross-national topics, and one third consists of in-depth, topical modules. Recent ISSP modules include the environment, gender and work, national identity, and the role of government. Recent topical modules include work organizations, multiculturalism, emotions, gender, mental health, giving/volunteering, altruism, Internet, and genetics. The data sets are available on the GSS Web site.

Research Opportunities

Several important types of research are facilitated by the GSS design. First, the replication of items allows the study of societal change. Moreover, because all surveys and all variables are organized in one cumulative file, researchers do not have to patch together time series from different and often incompatible data sets. By just running the data by YEAR, more than 1,600 trends can be tracked.

Second, replication also means that subgroups can be pooled across surveys to aggregate an adequate sample for analysis. For example, Blacks at about 12% of the population account for about 175 respondents in a 1,500 case sample—too few for detailed analysis. But in the 1972–2004 GSSs there are 6,399 Blacks—more than enough for analysis.

Third, researchers can both track trends and pool cases. For example, Blacks from the 1970s, 1980s, 1990s, and 2000s can be combined to have four time points and still have between 1,216 and 2,208 Blacks in each subsample.

Fourth, the 18 ISSP studies (1985–2005) offer the largest and most accessible body of cross-national social science data available. Moreover, reflecting the GSS's core interest in societal trends, the ISSPs have an across-time component. For example, the role-of-government topic in 1985 was repeated in 1990, 1996, and 2006. Thus, the GSS/ISSP has both a cross-national and across-time perspective.

Finally, the GSS's detailed and extensive set of demographics allows in-depth analysis of background influences. For example, the GSS does not have merely a single measure of education, but eight standard measures: the exact number of years of schooling completed and the highest degree obtained for respondents, mothers, fathers, and spouses. For occupation, the GSS has three-digit census codes, International Standard of Occupation Codes, NORC-GSS prestige scores, and Duncan Socioeconomic Index values for respondents, parents, and spouses.

Impact of the GSS

As the largest and longest-running project of NSF's Sociology Program, the GSS has had a tremendous impact on social science research. The GSS has been used in approximately 14,000 publications, and new usages accumulate at more than 700 per year. Among top sociology journals (*American Journal of Sociology*, *American Sociological Review*, and *Social Forces*), GSS use is second only to the U.S. Census.

The GSS has also had a large influence on college teaching. Millions of students have learned about society and research methodology in courses using the GSS. More than 400 textbooks in sociology, political science, statistics, and other fields utilize the GSS.

The GSS has aptly been described as a "national resource" by the National Academy of Science and as a "public utility for the community at large" (NSF).

The GSS is grounded in the past but growing into the future. It combines replication and innovation, incorporates both the societal change and comparative perspectives, and joins patrician quality standards with plebeian dissemination. Through these synergies it serves the social science communities and others.

Tom W. Smith

See also International Social Survey Programme (ISSP); National Opinion Research Center (NORC)

Further Readings

General Social Survey: http://www.gss.norc.org
International Social Survey Programme: http://www.issp.org

GEOGRAPHIC SCREENING

Most surveys target a specific geopolitical area, so that estimates produced from their data can be representative of that area. For some surveys, the area consists of an entire nation, but other surveys aim to produce regional estimates (such as those for states, counties, or zip codes). Thus, such surveys require some sort of geographic screening, or determination that a sampled case falls within the target geography, to establish study eligibility. If the screening is inherent in the sampling design itself, no further information is required. Other studies require additional screening steps, either prior to sample release or during the field period. Decisions about the level of geographic screening for a study arise from the sampling frame to be used.

When the sampling frame for a desired geographic area can be tied clearly to that area, no screening is needed beyond the design of the sample itself. For example, the sampling frame for a mail-based survey is composed of addresses that are known to be within a specific geographic area. Thus, geographic screening is part of the sampling design itself. Similarly, the sampling frame for an area probability sample is, by definition, geopolitically based, and therefore, no additional geographic screening is needed.

Telephone surveys typically use sampling frames that are defined by areas such as the nation as a whole, states, counties, cities, Census tracts, or zip codes. Samples of telephone numbers are generated by linking telephone exchanges to the desired target geography. In random-digit dialing (RDD) surveys of relatively small areas, it is impossible to match exactly telephone numbers with the boundaries of the target area. Researchers must determine whether the level of agreement between sampled telephone exchanges and the geography of interest is sufficient for their purposes or whether further questioning of the respondents to establish their location is warranted. This questioning can be complex and difficult to operationalize, thus leading to errors of omission and commission in which some eligible people are incorrectly screened out and some ineligible people are incorrectly screened in.

Implementation of additional screening steps increases the likelihood that all sample units are within the target geography, though it is likely to increase study costs as well—both of these are factors that influence a researcher's decision on the need for further screening. Decisions on additional screening are also influenced by considerations of incidence and coverage. Incidence is the degree to which the geography is represented in the sampled telephone exchanges, whereas coverage is the proportion of sampled telephone exchanges that are represented within the target geography. RDD sampling frames are often used to maximize coverage—ideally, all telephone numbers in an area are included in the frame for the RDD sample, including unlisted numbers, which would be missed if the sample were selected from telephone listings. However, RDD samples usually do not enjoy a perfect match between telephone exchanges and the target geography, so some sampled cases may lie outside the study boundaries. In such situations, the screening interview must ask respondents whether they reside in the target area. For example, respondents may be asked whether they live in a given county, zip code, or an area bounded by selected roads or other geographic markers.

If incidence is of greater concern than coverage, a list sample can be used. A list sample is based on residential, directory-listed telephone numbers that fall within a specified geographic area, thus ensuring the location of sampled cases prior to sample release, although such lists may contain inaccuracies due to people moving while keeping their same telephone number. (This issue is becoming more problematic in the United States since the inception of number portability in 2004.) List samples increase incidence but lower coverage, because unlisted telephone numbers are excluded. A similar method of geographic screening uses geo-coding to match addresses to all possible telephone numbers in an RDD sample and thus determine which sampled numbers are within the target geography. Then, only those cases that fall within the geography are retained in the sample. Geo-coding can improve the efficiency of an RDD sample, but it decreases coverage in the same way as a traditional list sample. Neither list samples nor geo-coded samples require respondents to answer geographic screening questions, because the sampling design ensures that cases are within the target boundaries. Though both have imperfect coverage, they can increase the cost-effectiveness and efficiency of data collection, particularly for studies targeting rare subgroups or small areas.

Web-based surveys have the same geographic screening limitations as telephone surveys. If the sampling frame for a Web-based survey can be tied to a geographic area during the design phase, it is not necessary to further screen respondents upon contact. However, if estimates by geography are desired and there is doubt about the respondent's residence, as often happens in Web-based studies, then geographic screening should be included in the questionnaire.

The increased use of cellular telephones (accompanied by portable telephone numbers) and Voice over Internet Protocol (VoIP) technology poses a geographic screening challenge. The likelihood that pre-screening procedures can adequately determine the geographic location of a sampled telephone number is substantially lower for cell phones and VoIP, because the usual telephone exchange identifiers are not necessarily tied to a specific geography in the way that they are for landline telephones. Thus, it is likely that all surveys with sampling frames that include cell phones or VoIP will require additional screening of respondents to ensure that they reside within the target geography.

Larry Osborn

See also Cell Phone Sampling; Errors of Omission; Errors of Commission; Number Portability; Representative Sample; Sampling Frame; Screening

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.

Marker, D. A., Waksberg, J., & Athey, L. (1990). Identifying geographic location from respondent interviews using RDD surveys. *Proceedings of the Section on Survey Research Methods* (pp. 500–504). Alexandria, VA: American Statistical Association.

Murphy, J., Burke, A., & Murphy, W. (2003). Assessing the validity of matched address data for unlisted telephone numbers in REACH 2010. *Proceedings of the Section on Survey Research Methods* (pp. 2972–2976). Alexandria, VA: American Statistical Association.

Sudman, S. (1985). Efficient screening methods for the sampling of geographically clustered special populations. *Journal of Marketing Research*, *22*(1), 20–29.

Gestalt Psychology

Often summarized by the phrase "The whole is greater than the sum of its parts," Gestalt psychology refers to an approach to understanding everyday human experiences as a whole rather than breaking them down into a collection of individual stimuli, behaviors, or both. This approach recognizes the ability of the human brain to piece together separate stimuli in context to one another and their surroundings so that the overall impression of an object, event, or other stimulus provides more information to the individual making the observation than was provided by the individual component stimuli. In other words, the individual may actually experience something that is not present in the stimuli themselves. A common example of this is watching a motion picture at a theater. Motion pictures (on film) actually consist of a series of still shots presented in rapid succession to give the impression of movement. Any one frame of the movie alone is simply a still photograph. When presented in rapid succession, however, the brain is able to fill in the gaps so that the individual has the experience of fluid motion.

This ability of the human brain, referred to as the *phi phenomena*, was used by Max Wertheimer to demonstrate the value of a holistic approach to studying psychology. Since that time, many other principles of Gestalt psychology have been identified. These include emergence, reification, multi-stability, and invariance. *Emergence* occurs whenever there is confusion between figure and ground in an image. The figure of an image refers to the subject or object, whereas the ground refers to the setting or background. The classic example of emergence in psychology texts is a black and white picture that initially appears to be random splotches of black ink (figure) on a white paper (ground). When the individual trains his or her eye on the white portion of the picture as the figure instead of the ground, a picture of a spotted Dalmatian dog appears.

Reification is similar to emergence in that the phenomenon is based on the visual relationship between figure and ground. Reification, however, is more often associated with the arrangement of geometric shapes, whereby the relationship of the shapes (figures) on the ground begin to form a shape of the ground. Hence, the ground becomes the figure. *Multi-stability* refers to the tendency for an ambiguous figure to be interpreted as two or more different figures such that the brain cannot decide which figure is correct. This phenomenon can be isolated to the figure itself (e.g., Necker's cube), as well as a product of figure/ground confusion (e.g., Rubin's Figure/Vase Illusion). Finally, the principle of *invariance* refers to the brain's ability to recognize simple objects, regardless of distortion (e.g., size, position/rotation).

More important to survey research are the laws of prägnanz, similarity, proximity, and closure. Each of these is critical to the design and layout of self-administered questionnaires, whether on paper or online. The *law of prägnanz* ("good form") is defined by the observation that individuals tend to group stimuli in a way that is most easily understood. In other words, the human brain attempts to interpret stimuli in the simplest and most orderly way possible. Thus highly complex graphics and symbols, or unnecessary variations in the layout of questions may increase cognitive burden on the respondent and reduce the quality and quantity of survey responses. The tendency to group similar items is another way the brain attempts to simplify visual elements. This tendency is known as the *law of similarity*. This again emphasizes the importance of consistent labeling of sections of a questionnaire. Good labels provide structure for the respondent to aide them in completing the questionnaire.

The *law of proximity* states that elements that are closer together, either physically or temporally, are more likely to be cognitively grouped than those farther apart. Thus response options should be evenly spaced so that they are mentally grouped and attention is not drawn to any one response or set of responses unless that is desired explicitly. Finally, the brain may add elements to a figure to more easily understand it as a whole. For example, a dotted line is perceived as a line rather than a series of dots. This is referred to as the *law of closure*, as the brain attempts to close the gaps in the image.

In conclusion, it is important to be aware of figure/ground relationships when laying out a questionnaire. In fine art this is commonly referred to as being aware of the "negative space" on a page. It is possible to communicate unintended messages to a respondent by focusing solely on the figure and ignoring the ground when creating an image. Likewise, in designing a survey instrument, researchers must be aware of the spacing of graphics and symbols so that unintended connections of closures are minimized.

Kenneth W. Steve

See also Cognitive Aspects of Survey Methodology (CASM); Graphical Language; Questionnaire Design; Respondent Burden; Visual Communication; Web Survey

Further Readings

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.
Hilgard, E. R. (1987). *Psychology in America: A historical survey*. New York: Harcourt Brace Jovanovich.
Kendler, H. (1987). *Historical foundations of modern psychology*. Pacific Grove, CA: Brooks/Cole.

Graphical Language

Respondents interpret the meaning of survey questions from both the verbal and graphical language used in the questionnaire. Graphical language includes various elements such as contours and lines, images, numbers, and symbols and their attributes such as movement, spatial location, color or contrast, and size. These graphical elements influence how respondents perceive survey information and therefore significantly impact the survey response process. Graphical language can convey meaning independently, or it can influence or modify how written text is perceived. Thus, it can be compared to paralanguage that is conveyed aurally through a speaker's voice (e.g., inflection, tone) and to nonverbal communication in face-to-face interactions (e.g., gaze, facial expressions, body language, and gestures). Because paper and Web surveys transmit information visually, survey designers can strategically use graphical language to convey information and meaning to respondents. However, graphical language may also confuse survey respondents when used in competing ways, carelessly, or inconsistently.

Graphical language acts like a visual paralanguage to emphasize or draw attention to information in a survey, create groupings and subgroupings of information, and improve navigation through the survey. Graphical attributes such as size, contrast, color, layout, and position can influence the meaning assigned to written text in many ways. For example, in Figure 1, the larger size and use of reverse print for the question number and the "Next Page" button and the underlining of the word *satisfaction* in the question stem help draw respondents' attention to this information. In addition, locating the question number "1" in the upper left of the screen helps convey to respondents that the number, one, means this is where they should begin.

Furthermore, graphical language can encourage respondents to perceive information as belonging together in a group and therefore as related conceptually. The Gestalt principles of proximity, similarity, connectedness, and common region indicate that information is grouped visually when items are located near each other, share similar graphical attributes (shape, size, color/contrast, etc.), and are connected or enclosed within a common region such as a square. For example, in Figure 1, using similar size, font, and reverse print for the question number and "Next Page" button encourages respondents to group them visually and then conceptually as tools to aid in navigating through the survey. In addition, using a larger size for the question stem but similar font size for each item helps respondents perceive the subgroups within the question group (i.e., response items separate from question stem). Grouping is also established by the gray lines in Figure 1 that connect the text of each item to the appropriate answer spaces and by positioning the radio buttons in closer proximity horizontally than vertically.

In addition to acting like a visual paralanguage, graphical elements such as symbols, logos, pictures, and other images can independently influence the tone of printed survey contacts (letters or emails), instructions to respondents, individual questions, and response categories. Appropriate logos and images on contact letters and survey instruments can increase respondent motivation and commitment to completing the survey. Moreover, pictures and other images can be used to convey information or enhance the meaning of written text in much the same way that facial expressions, body language, and gestures do in face-to-face communication. For example, in Figure 1 the combination of facial expression images and numbers are used to convey the meaning of each scale point.

Since research on Web surveys has shown that pictures and other graphical images can modify the meaning respondents assign to particular questions and concepts, images must be chosen carefully to avoid negative impacts on measurement, such as when the inclusion of pictures of an endangered species artificially increased respondent support for that species. Moreover, research has shown that including sizable graphical elements in Web surveys can slow page download times, thus increasing respondent burden and sometimes nonresponse. The increased use of Web surveys has heightened the attention given to graphical language in survey questionnaire design because graphical language is easy and inexpensive to include and modify in Web surveys (i.e., no printing costs). In addition to increasing the need for research into the effects of

Figure 1 Examples of the use of graphical language in a Web survey

specific graphical language elements on survey response, the greater use of graphical language has also created the need for enhanced Web programming to maintain visual consistency and standardize the survey stimulus across different configurations of respondent hardware, software, and preferences.

Leah Melani Christian and Jolene D. Smyth

See also Gestalt Psychology; Measurement Error; Questionnaire Design; Radio Buttons; Visual Communication; Web Survey

Further Readings

Redline, C. D., & Dillman D. A. (2002). The influence of alternative visual designs on respondents' performance with branching instructions in self-administered questionnaires. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 179–193). New York: Wiley.

Tourangeau, R., Couper M. P., & Conrad, F. (2004) Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368–393.

Guttman Scale

Given a data set of a sample of N persons and of a selection of n survey questions (variables) designed for measuring a particular trait—such as people's position on a political issue or their ability in a specific field of human activity—a Guttman Scale is the *hypothesis* that the data set would have a cumulative structure, in the following sense: For *any* two persons in the observed sample, one of them would exhibit all the manifestations of the trait that the other person would, and possibly additional ones. That is, there would be no two persons in the sample with the one person higher than the other in one variable but lower than the other in another variable.

To the extent a Guttman Scale hypothesis is supported in a data set, it can be said that all observed persons (and all their observed profiles) are comparable with respect to the selection of observed variables. And—if the data set is sufficiently rich—it can be inferred that members in the population represented by the sample can be meaningfully measured with respect to the trait (represented by the selection of variables) using a single (i.e., one-dimensional) scale.

Example

Consider, for example, public attitudes toward intervention in foreign countries as a trait in question. Presenting an appropriate sample of the adult population with the following questions can serve in measuring this attitude:

> *In the interest of our national security, do you favor spending money for encouraging a change in the policies of foreign country (C)? 1. No; 2. Yes, but only through nongovernmental organization (tax-deductible) funds; 3. Yes, any money, including nationally budgeted (taxpayers') money.*
>
> *In the interest of our national security, do you favor sending special civilian advisors in order to encourage a change in the policies of country (C)? 1. No; 2. Yes, but only as part of the normal diplomatic mission; 3. Yes, any size of special civilian mission, as necessary.*
>
> *In the interest of our national security, do you favor sending military advisers abroad in order to encourage a change in the policies of country (C)? 1. No; 2. Yes.*
>
> *In the interest of national security, do you favor sending our military troops abroad in order to encourage a change in the policies of country (C)? 1. No; 2. Yes.*

A respondent who, for the said purpose, favors spending national funds (score 3), sending any size of civilian mission (3) as well as military advisors (2) but objects to sending troops (1), would have the profile 3321, which is comparable to, and represents a more positive intervention attitude than, profile 3221, because the former is equal to or greater than the latter on every variable. But 2111 would be said to be *incomparable* to 1232, because at least on one variable the former is higher than the latter, and at least on one variable the former is lower than the latter. If, in a particular survey, all profiles are comparable—that is, no incomparable pair of profiles is observed to occur—the result is a Guttman Scale. In the previously mentioned example, if out of the $\times 3 \times 2 \times 2 = 36$ technically possible profiles, only the 7 profiles shown in the first column of Table 1 are actually observed, the data would constitute a Guttman Scale (or a *cumulative scale*).

Table 1 Example of a Guttman Scale with a single score assigned to observed profiles

Profile Based on Observed Responses	*Score* x *Assigned to the Profile*
1111	1
2111	2
3111	3
3211	4
3221	5
3321	6
3322	7

In this list of profiles each profile represents, in a strict sense, a more positive attitude toward intervention than its predecessor. Hence a single score x, from any set of ordered numbers, can be assigned to every observed profile such that (a) the higher x is, the more positive is the attitude toward intervention; and (b) given this assignment, a respondent's answers to all questions can be reproduced from his or her single score, x. Thus, if a Guttman Scale holds, the original four dimensions suggested by the four variables (and, in general, any number of variables, n) are reduced to a one-dimensional measurement scale.

Note that the order in which the variables are considered (and hence the order of their respective scores within the profiles) is immaterial for the definition and analysis of a Guttman Scale and so may be chosen as convenient. Also note that no weights are assumed to be associated with the variables.

When Data Do Not (Guttman) Scale

While pure Guttman Scales are rarely found in social research, approximate Guttman Scales have been found, for example, in studies of people's knowledge in specific areas, of the possession of household appliances, and more. Different versions of a *coefficient of reproducibility* have been proposed for assessing the degree of fit of data to a Guttman Scale. It is important to realize that deviations from a Guttman Scale can be of two kinds: (1) random deviations, suggesting the existence of a Guttman Scale with "noise," and (2) structured deviations, suggesting that two or more scales are needed to measure the studied trait meaningfully. Developing the procedures of multiple scaling, using partial order scalogram analysis by base coordinates (POSAC), Samuel Shye has generalized Guttman Scale to dimensionalities higher than one.

Misconceptions

Several misconceptions have accompanied the notion of the Guttman Scale throughout the years:

1. Some have sought to "construct" a Guttman Scale by eliminating variables (or respondents) from their data. Assuming one has had a rationale for selecting variables as representing a concept (and for defining the sampled population), such eliminations in a particular application may be questionable and should be avoided, except possibly in the context of the larger cycle of scientific investigation where concepts are reshaped and redefined. As noted, a Guttman Scale is essentially a hypothesis, which may or may not be supported by data.

2. Many believe that a Guttman Scale necessarily involves only dichotomous variables. Indeed, most illustrations and many applications in the literature are with such variables. As the example presented earlier shows, this need not be the case. However, when a Guttman Scale is found in dichotomous variables, the variables are naturally ordered according to their sensitivity in detecting the presence of the measured trait.

3. Confusion often arises between the Guttman Scale, which completely orders respondents on a one-dimensional scale, and the *simplex*, which, in Guttman's work, means a simple ordering of the variables on a one-dimensional line by an aspect of their contents. Such ordering of the variables is tested for by Faceted Smallest Space Analysis, a multidimensional scaling procedure that maps variables according to a similarity measure (typically correlation coefficient) between them. The mathematical relationship between the two kinds of spaces—that is, the measurement space into which respondents are mapped and the trait concept-space into which variables are mapped—has been studied within the theory of multiple scaling by POSAC.

As a procedure that ties substantive contents with logical aspects of data, the Guttman Scale heralded the development of Facet Theory by Louis Guttman and his associates.

Samuel Shye

Further Readings

Coombs, C. H., Coombs, L. C., & Lingoes, J. C. (1978). Stochastic cumulative scales. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences*. San Francisco: Jossey-Bass.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, *70*, 755–768.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Shye, S. (1985). *Multiple scaling: The theory and application of partial order scalogram analysis*. Amsterdam: North-Holland.

Shye, S. (1998). Modern facet theory: Content design and measurement in behavioral research. *European Journal of Psychological Assessment*, *14*(2), 160–171.

Shye, S., & Elizur, D. (1994). *Introduction to facet theory: Content design and intrinsic data analysis in behavioral research*. Thousand Oaks, CA: Sage.

H

Hagan and Collier Selection Method

The ideal method of selecting respondents within households is a probability method, but researchers seldom work under ideal conditions. They usually seek to improve within-unit coverage without adding to nonresponse. Probability selection methods, such as the Kish method, may enhance within-unit coverage but also may increase nonresponse because of the need for longer introductions, which are potentially intrusive and may alienate respondents. Longer surveys add to interviewer costs. Hagan and Collier developed a nonprobability quota technique that they believed would secure respondent cooperation more readily because it asked no household composition questions. Their procedure was based on the Troldahl-Carter-Bryant (T-C-B) respondent selection method, which in turn was based on the Kish technique.

Interviewers are given four forms, which are randomly distributed. The result is an oversampling of male respondents (desirable because of higher noncontact-related nonresponse among men). Within the eligible age range for a given survey, Form A requests the youngest male in the household; Form B asks for the oldest male; Form C, the youngest female; and Form D, the oldest female. If no such person resides there, the interviewer is instructed to ask for the person in the household of the opposite sex and same age grouping. Forms A, B, and C are each used two times in seven, and Form D is used one time in seven. Hagan and Collier reported favorable results in a sample compared with one that selected respondents by the T-C-B procedure. Demographic characteristics were similar, and the refusal rate at respondent selection was almost 5% less than the T-C-B method. Both methods have a small within-unit coverage bias because adults in households of more than two adults of the same sex whose ages are between the oldest and youngest adults have no chance of selection. Also, in three-adult households, one of the three adults would have the chance of being designated the respondent twice. Troldahl and Carter considered these violations of random sampling to be very small. Research using census data has shown that the bias caused by the Hagan-Collier method is very slight.

An example of Hagan-Collier question wording is *May I please speak to the "youngest man"?* Another example is *For this survey, I need to speak with the youngest adult male in your household over the age of 17, if there is one*. If there is none, the following question is asked: *Then may I please speak with the youngest adult female?* Wording should include the fact that the designated respondent is not the one who happens to be at home at the time but, instead, is the one who lives in the household. Interviewers need training in awareness that a woman in a one-person household fits as either the youngest woman or the oldest woman, that “youngest man” can apply to an elderly male, and that informants can be confused and think the interviewer is asking for an old man (or a young woman), among other things.

Cecilie Gaziano

See also Coverage Error; Kish Selection Method; Troldahl-Carter-Bryant Respondent Selection Method; Within-Unit Coverage; Within-Unit Selection

Further Readings

Bryant, B. E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research*, *12*(May), 129–135.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Hagan, D. E., & Collier, C. M. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly*, *47*, 547–556.

Krótki, K. P., & Porcellini, L. (1995). Respondent selection bias in the Hagen [sic] -Collier approach. *Proceedings of the Survey Research Methods Section* (pp. 694–696). Alexandria, VA: American Statistical Association.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Troldahl, V., & Carter, R. E., Jr. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, *1*(May), 71–76.

HALF-OPEN INTERVAL

The half-open interval is a linking procedure that is used in some surveys to address issues of noncoverage. Sampling frames or lists are not perfect, and survey researchers often use frames with problems such as missing elementary units, blanks for relevant information, clusters of elementary units, and duplicate listings. Of these problems, sample frames that are missing elementary units—known as *noncoverage*—frequently present important practical problems. For example, housing lists incorporating addresses are often used in household surveys. The housing list is often out of date, and when an interviewer visits the housing unit selected from the list, there can be newly constructed housing units that were not on the original list used for sampling. When there is noncoverage of the target population due to an imperfect sample frame, specific remedies are required to improve the frame coverage.

To account for the units missing from a frame, researchers may use a linking procedure. This is a useful device in many situations where the missing units are scattered individually or in small clusters. The linking procedure is often called the *half-open interval*, which indicates the interval by the boundary between the selected unit and the next unit in the list.

For example, suppose that 100 Kish Street has been selected from the housing list. From an area frame perspective, the address of 100 Kish Street has a boundary defined by its property lines up to the property of the next address on the housing list, 102 Kish Street. This boundary denotes the half-open interval, which does not include the next address on the list. If there is a new or missed housing unit within the interval (e.g., at 100 1/2 Kish Street), an interviewer treats it as a sampled unit and conducts an interview with all the units in the interval, including the pre-specified unit (i.e., 100 Kish Street). Thus, the missed units have the same probability of selection as the pre-specified unit.

Occasionally, ordinary linking procedures cannot deal with the missed units adequately. For example, in the case of an address where a single household is expected, finding a newly built apartment house with 20 dwellings presents a real dilemma, since the interviewer technically would need to conduct 21 interviews instead of 1 interview. In such cases, the additional dwellings may be subsampled to reduce the interviewer workload, and weighting must be implemented to compensate for any unequal probabilities of selection. Instead of linking procedures, a large number of newly constructed units can be put into a supplementary stratum from which they are selected with varying probabilities, although it is better if a check of the frame is available to avoid such unpleasant occurrences.

These linking procedures can also be applied to the instances where ordered lists serve as frames. For example, on a payroll listing, a new employee in a department can be missed if the frame is slightly out of date. If a unique employment position on the list, such as the employee listed last in each department, is selected as a sample unit, then the new employee is also added to the frame and selected with the same probability as the one listed last by the linking rule. Another example is the list a public school has of its pupils' households. When the household of a selected child is visited, recently born or missed children can be discovered. The linking procedure must fit the missed children into the half-open interval, thereby reducing the noncoverage of the original list.

SunWoong Kim

See also Noncoverage

Further Readings

Innacchione, V. G., Staab, J. M., & Redden, D. T. (2003). Evaluating the use of residential mailing addresses in a metropolitan household survey. *Public Opinion Quarterly*, *67*, 202–210.

Hang-Up During Introduction (HUDI)

A telephone interview that is terminated by the respondent during the introduction of the interview shortly after an interviewer has made contact is called a *hang-up during introduction* (HUDI). HUDI is a form of refusal to the survey request that is growing in occurrence and provides little or no opportunity for the interviewer to overcome the respondent objection.

The most difficult time to assure success in a telephone interview is during the first few seconds of the call. During this time, the interviewer has to identify the purpose and legitimacy of the call. In the past two decades, there has been an increasing tendency for respondents to hang up on the interviewer during this time without completing a full interaction with the interviewer. In contrast, in the 1970s and 1980s when telephone surveys were first gaining their legitimacy as a valid survey method of the public, there was a social norm that held most people to not hang up on a stranger (the interviewer who called them) abruptly. However, with the problems causes by excessive telemarketing in the 1990s and busy lifestyles, people are far less reluctant to just hang up.

Early work in the late 1980s found that 40% of refusals occur in the first two sentences of the introduction. Similarly, more recent research has found that HUDIs last an average of 15 seconds. A study in 2003 found that one in four HUDIs occur without the respondent saying anything at all to the interviewer, and a 2005 study found two fifths of respondents hanging up on the interviewer without speaking. Respondents may give brief and abrupt objections, which are most frequently an indication of "not interested" or "don't have time" and then abruptly hang up.

Urbanicity has been found to be negatively associated with response rate. This finding is reflected in the incidence of HUDIs by metropolitan area size. A 2005 study found a 6 percentage point gap in the occurrence of HUDIs in the 10 largest metropolitan areas compared to cities and towns of less than 200,000 population. The Northeast and West regions of the United States show the highest rates of HUDIs while the Midwest rate was found to be 5 points lower. The study also showed that households that had been sent a pre-notification mailer (advance letter or postcard) were less likely to hang up during the introduction.

The interviewer is challenged with trying to establish rapport and engage the respondent while introducing the purpose of the call. In the 1970s and 1980s, interviewer research focused on identifying the words to use in a scripted introduction to improve respondent cooperation rates. The result was the identification of preferred words but no clear agreement on the benefit of scripted introductions. During the 1990s, attention shifted to techniques of allowing interviewers to tailor introductions to engage the respondent. The goal is to maintain interaction with the respondent. Studies have found that the longer interaction is maintained, the more likely it will result in a completed interview.

The success of maintaining interaction is dependent on the interviewer skill as well as the words of the introduction and the respondent behavior. More recent research has shifted from identifying the specific words to say to identifying the interviewer characteristics that best predict a successful interview. However, with HUDIs, the interviewer has no time to respond to an objection, if one is voiced. There have been renewed efforts to develop interviewer refusal aversion training to prepare interviewers to quickly identify the most important concern of the respondent and how to overcome the objection. In addition, research is exploring characteristics of successful interviewers, including voice characteristics, which result in higher respondent cooperation rates.

There is much work to be done to better understand the characteristics of HUDIs, the correlates of incidence, and how to combat this increasing challenge to successful telephone survey interviewing.

Barbara C. O'Hare

See also Advance Letter; Interviewer Training; Introduction; Refusal; Refusal Avoidance; Respondent Refusal; Tailoring

Further Readings

Couper, M. P., & Groves, R. M. (2002). Introductory interactions in telephone surveys and nonresponse.

In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 161–178). New York: Wiley.
Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.
Mooney, C., & O'Hare, B. C. (2006, May). *HUDIs: A look at quick refusals*. Paper presented at the American Association of Public Opinion Research, Montreal, Quebec, Canada.
Oksenberg, L., & Cannell, C. F. (1988). Effects of interviewer vocal characteristics on nonresponse. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 257–272). New York: Wiley.

Hansen, Morris (1910–1990)

Morris Hansen was one of the most innovative and influential statisticians of the 20th century. He helped pioneer the work of sampling techniques and the use of the total survey error perspective in designing surveys. He also developed quality control procedures for surveys that helped legitimize the accuracy of survey research. He attended the University of Wyoming (B.S.) and American University (M.S.) and had a long career at the U.S. Census Bureau and Westat, until his death in 1990.

Morris Hansen was born in Thermopolis, Wyoming, in 1910 and spent his formative years in Worland, Wyoming. He earned a B.S. in accounting from the University of Wyoming in 1934. After graduation he started his career at the U.S. Census Bureau in Washington, D.C. Fascinated by statistics in college, Morris started his formal training as a statistician when he arrived in Washington, taking evening courses at the Graduate School of the U.S. Department of Agriculture and eventually earning a master's degree in statistics from American University in 1940.

During his first years at the U.S. Census Bureau, Morris started to establish himself as a highly skilled statistician. At age 26, Morris worked on the sample design for an unemployment survey (which would later evolve into the Current Population Survey) for the federal government, an innovative project because, at the time, the government preferred using census data for their studies. Hansen convinced them that data from a survey could be as reliable, or even more so, as data collected from a census. To accomplish this, Morris introduced the concept of total survey error, which takes into account all of the sources of error in a survey, for example, interviewer effects, questionnaire design, and so forth. The cost savings that result from collecting data from a sample of the population, instead of the entire population, can then be spent on reducing the error from other sources. Although sampling introduces some error into the data, the total survey error is reduced because of a reduction of error from a multitude of sources. Hansen, along with William Hurwitz, would further develop the mathematical theory that underlies sampling methodology.

Hansen retired from the U.S. Census Bureau in 1968. Shortly thereafter he was invited to join Westat, which was, at the time, a small research organization. Morris accepted the invitation and joined Westat as a senior vice president. He would later serve as chairman of the board, after Ed Bryant, a founder of Westat and the preceding chairman, retired. While at Westat, Morris led many important government projects, such as the Consumer Price Index and the National Assessment of Educational Progress. He also developed new techniques for quality control in survey research. Morris did not retire from Westat and continued to work vigorously on statistical and survey methods until his death in 1990. Morris is remembered as an inspiration, a great collaborator, and a passionate teacher by those who worked with and learned from him.

Paul Schroeder

See also Current Population Survey (CPS); Total Survey Error (TSE)

Further Readings

Hansen, M., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematics Statistics*, *14*(4), 333–362.
Hansen, M., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory* (Vols. 1 & 2). New York: Wiley.
Hansen, M., Hurwitz, W. N., Marks, E. S., & Mauldin, W. P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, *46*, 147–190.
Hurwitz, M., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability sampling

inferences in sample surveys. *Journal of the American Statistical Association*, *78*, 776–793.

Hit Rate

In general, *hit rate* is a ratio or a proportion, and the term is used in many environments and disciplines with specific definitions for both the denominator and numerator. In the online world, it usually means the number of "hits" a Web page receives during some period of time. In marketing, it can mean the number of sales achieved as a percentage of the number of sales calls made. In survey research, hit rate is most commonly used to refer to the proportion of telephone numbers in a sample that are working residential numbers. However, hit rate is sometimes used to mean incidence. Both meanings of hit rate are essential components of sample size calculations.

In its common usage, hit rate is synonymous with the terms *working residential rate* and *working phone rate*. For a residential telephone sample, eligible units would be those numbers that connect to a household, while ineligible numbers would include nonworking or disconnected numbers, data/fax lines or numbers that connect to an ineligible unit such as a business. For an in-person survey it might mean the proportion of occupied housing units in the frame, and for a mail survey it could mean the proportion of deliverable mail pieces in the list.

Hit rate is also sometimes used as a surrogate for incidence or the proportion of qualified contacts to all contacts. For example, a survey might require screening households for further eligibility, such as living within a particular geography, having a certain income, or belonging to a specific racial or ethnic group. In these cases the hit rate would be the probability of finding members of that target population among all contacts.

Understanding and being able to estimate these two hit rates is integral to sample design. Most formulas for calculating the number of sample units needed to complete a set number of interviews include both of these definitions of hit rate (working phone rate and incidence) in conjunction with estimates of contact rate and completion rate for the survey.

Linda Piekarski

See also Contact Rate; Eligibility; Nonresidential; Out of Sample; Target Population

Horse Race Journalism

Contemporary political reporting, especially news that has a focus on elections and policy debates, is often covered as though these matters are a game among competing candidates and elites. Thus, this dominant approach to covering elections has come to be referred to by academics and others as *horse race journalism*, the *game schema*, or the *strategy frame*. Rather than foregrounding issue positions, candidate qualifications, or policy proposals, journalists instead tend to cast these features of the political terrain as secondary to a focus on who's ahead and who's behind in winning the campaign or a policy battle, the principal players (i.e., the generals and lieutenants) involved, and the shifting gamesmanship strategies and tactics employed.

Horse race journalism focuses almost exclusively on which candidates or players are most adept at gaining power while also undermining the political chances of opponents. A horse race is an apt metaphor, as this style of reporting translates easily into the conventions of sports coverage, with a focus on competing political gladiators who survive to campaign another day or who are the first to cross the finish line. Polling and public opinion surveys are a central feature of this political spectacle. In fact, it is polls and other surveys that supply most of the objective data for reporters to define who is winning while also providing news pegs for transitioning into attributions about the reasons for political successes and political failures.

The Dominance of Horse Race Journalism

Over the past 40 years, the rise in horse race journalism has been called by Thomas Patterson the "quiet revolution" in U.S. election reporting. Patterson's now classic analysis finds that coverage focusing on the game schema that frames elections in terms of strategy and political success rose from 45% of news stories sampled in 1960 to more than 80% of stories in 1982. In comparison, coverage focusing on policy schema, which frame elections in terms of policy and leadership, dropped from 50% of coverage in 1960 to just 10% of coverage analyzed in 1992.

Other analyses confirm the contemporary dominance of the horse race interpretation in election coverage. In one study of the 2000 U.S. presidential campaign, strategy coverage accounted for more than 70% of the TV stories at the major news networks. The most recent available analysis conducted by the Pew Center—tracking the first 5 months of 2008 presidential primary coverage—found that horse race reporting accounted for 63% of print and TV stories analyzed compared to just 15% of coverage that focused on ideas and policy proposals and just 1% of stories that focused on the track records or past public performance of candidates.

In the United States, not only has horse race strategy come to define elections, but the approach also increasingly characterizes more simplified coverage of what were originally considered complex and technical policy debates. First observed by Joseph Capella and Kathleen Jamieson in their analysis of the early 1990s debate over health care reform, when coverage of policy debates shifted from specialty news beats to the political pages, the strategy frame has been tracked as the dominant narrative in reporting of issues as diverse as stem cell research, climate change, food biotechnology, the Human Genome Project, and the teaching of evolution in schools.

Forces Behind the Quiet Revolution

Horse race journalism is fueled partly by industry trends and organizational imperatives, but also by journalistic need for narrative. In a nonstop and highly competitive news environment with a 24-hour news cycle and tight budgets, reporting the complexity of elections and policy debates in terms of the strategic game is simply easier, more efficient, and considered better business practice.

Public opinion surveys are a competitive advantage in the news marketplace; they are even an important part of media organization branding and marketing. Perhaps more important, polls help fill the demand for anything new in a day-long coverage cycle while also fitting with trends toward secondhand rather than primary reporting. The growth in the survey industry, and the almost constant release of new polling data, has helped fuel the rise in horse race coverage. For example, in analyzing trial heat polls tracking the presidential nominees, Michael Traugott reported a 900% increase in such polls from 1984 to 2000. In 2004, the total number of trial heat polls remained equivalent to that of the 2000 presidential campaign, but there was more of a mix of different types of polls, as several organizations focused specifically on anticipated battleground states. And, the increased use of tracking polls likely magnifies horse race coverage. In combination with economic imperatives and the increased availability of polling, horse race coverage also resonates strongly with the informal rules of political reporting. American journalists pay heavy attention to scandals, corruption, or false and deceptive claims, but because of their preferred objectivity norm, they typically shy away in their news stories from actively assessing whether one side in an election or policy debate has the better set of candidates, ideas, or proposed solutions. With a preference for partisan neutrality, it is much easier for journalists to default to the strategic game interpretation. Issue positions and policy debates are part of this coverage but very much secondary to a dominant narrative of politics that turns on conflict, advancement, and personal ambition.

Tom Rosenstiel has connected the objectivity norm to the new *synthetic journalism*, a trend that further favors poll-driven horse race coverage. In a hypercompetitive 24-hour news cycle, there is increasing demand for journalists to try to synthesize into their own coverage what already has been reported by other news organizations. This new information might include newly revealed insider strategy, the latest negative attack, or a perceived embarrassing gaffe or mistake. Such details, however, are problematic because the need to synthesize critical or damaging information runs up against the preferred norm of objectivity while also providing fodder for claims of liberal bias.

Yet, as news pegs, polls serve an important insulating function, providing journalists the “objective” organizing device by which to comment on and analyze news that previously has been reported by other outlets. For example, if a new survey indicates that a candidate is slipping in public popularity, the reporting of the poll’s results provides the subsequent opening for journalists to then attribute the opinion shift to a recent negative ad, character flaw, allegation, or political slipup. Kathleen Frankovic has noted a dramatic rise not only in the reporting of specific poll results but also in rhetorical references to “polls say” or “polls show,” with close to 9,000 such general mentions in her sample of newspapers in 2004 compared to roughly 3,000 such mentions in 1992. This reliance on “the authority of polls” adds perceived precision and objectivity to journalists’ coverage. According to Frankovic, this rhetorical innovation allows journalists to make independent attributions about candidate success or failure without relying on the consensus of experts. Moreover, she argues that the heightened emphasis on the polls alters the criteria by which audiences think about the candidates, shifting from a focus on issue positions and qualifications to that of electability.

Of course, an accent on strategy, ambition, poll position, and insider intrigue is not the only way that political reporters can translate an election campaign or policy debate for audiences. Journalists, for example, could alternatively emphasize issue positions; the choice between distinct sets of ideas and ideologies; the context for policy proposals; or the credentials and governing record of candidates and parties. Yet, in comparison to the horse race, the storytelling potential of each of these alternative ways of defining what is newsworthy in politics is perceived as more limited. In fact, according to the norms that dominate most political news beats, once the issue positions, credentials, background, or track record of a candidate is first covered, they are quickly considered old news.

Reasons for Concern About Horse Race Journalism

Scholars have raised multiple concerns about the impacts of horse race journalism. Patterson and others fear that the focus on the game over substance undermines the ability of citizens to learn from coverage and to reach informed decisions in elections or about policy debates. Capella and Jamieson argue that the strategy frame portrays candidates and elected officials as self-interested and poll-driven opportunists, a portrayal that they show promotes cynicism and distrust among audiences. Stephen Farnsworth and Robert Lichter go so far as to suggest that horse race coverage in the primary elections results in a self-reinforcing bandwagon effect, with positive horse race coverage improving a candidate's standing in subsequent polls and negative horse race coverage hurting a candidate's poll standings. Their observation fits with what many political commentators and candidates complain about: that overreliance on polling narrows news attention and emphasis to just the two to three leading candidates while overemphasizing perceived electability as a criterion for voters to consider. In this sense, horse race coverage can be perceived as unduly promoting the media as a central institution in deciding electoral outcomes. In terms of horse race coverage of policy debates, other than failing to provide context and background for audiences, Matthew Nisbet and Michael Huge argue that the strategy frame's preferred "he said, she said" style leads to a false balance in the treatment of technical issues where there is clear expert consensus. Polling experts offer other reservations. For example, Frankovic and others warn that overreliance on horse race journalism and polling potentially undermines public trust in the accuracy and validity of polling.

Matthew C. Nisbet

See also Bandwagon and Underdog Effects; Polls; Pollster; Precision Journalism; Public Opinion; Public Opinion Research; Tracking Polls; World Association for Public Opinion Research (WAPOR)

Further Readings

Capella, J. N., & Jamieson, K. H. (1997). *Spiral of cynicism: The press and the public good*. New York: Oxford University Press.

Farnsworth, S. J., & Lichter, S. R. (2003). *The nightly news nightmare: Network television's coverage of US presidential elections, 1988–2000*. Lanham, MD: Rowman & Littlefield.

Frankovic, K. A. (2005). Reporting "the polls" in 2004. *Public Opinion Quarterly*, *69*, 682–697.

Nisbet, M. C., & Huge, M. (2006). Attention cycles and frames in the plant biotechnology debate: Managing power and participation through the press/policy connection. *Harvard International Journal of Press/Politics*, *11*(2), 3–40.

Patterson, T. E. (1977). The 1976 horserace. *Wilson Quarterly*, *1*, 73–79.

Patterson, T. E. (1993). *Out of order*. New York: Knopf.

Patterson, T. E. (2005). Of polls, mountains: U.S. journalists and their use of election surveys. *Public Opinion Quarterly*, *69*(5), 716–724.

Pew Project for Excellence in Journalism. (2007, October 29). *The invisible primary—invisible no longer: A first look at the 2008 presidential campaign* [Press release and report].

Rosenstiel, T. (2005). Political polling and the new media culture: A case of more being less. *Public Opinion Quarterly*, *69*(5), 698–715.

Traugott, M. (2005). The accuracy of the national preelection polls in the 2004 presidential election. *Public Opinion Quarterly*, *69*(5), 642–654.

Hot-Deck Imputation

Hot-deck imputation is a popular and widely used imputation method to handle missing data. The method involves filling in missing data on variables of interest from nonrespondents (or recipients) using observed values from respondents (i.e., donors) within the same survey data set. Hot-deck imputation can be applied to missing data caused by either failure to participate in a survey (i.e., unit nonresponse) or failure to respond to certain survey questions (i.e., item nonresponse). The term *hot deck*, in contrast with *cold deck*, dates back to the storage of data on punch cards. It indicates that the donors and the recipients are from the same data set; the stack of cards was "hot" because it was currently being processed (i.e., run through the card reader quickly, which heated the punch cards). Cold-deck imputation, by contrast, selects donors from external data sets.

This entry describes the various types of hot-deck imputation: sequential, hierarchical, and nearest neighbor. This entry then discusses the assumptions underlying these methods and reviews the advantages and disadvantages of hot-deck imputation.

Sequential Hot-Deck Imputation

The basic idea behind hot-deck imputation is to match a recipient to a donor with similar characteristics and then transfer the donor's value to the recipient. There are various methods to match a recipient to a donor. The traditional hot-deck procedures begin with the specification of imputation classes constructed with auxiliary variables that are observed or known for both respondents and nonrespondents. Within each imputation class, the first nonmissing value (or record) is assigned as the potential donor. Each subsequent record is then compared to that potential donor; if the record has a nonmissing value, it replaces the potential donor. But if the record has a missing value, the most recent donor value is filled in. This is also called *sequential hot-deck imputation*.

A simple example explains this procedure. Given a sample of respondents and nonrespondents, the values on variable y are either observed or missing. If gender is known for all respondents and nonrespondents, two imputation classes can be constructed. The sequential hot-deck imputation procedure continually stores and replaces potential donor values from each nonmissing record. If a missing value on the y variable is found, the most recent donor value is then transferred to that nonrespondent.

The sequential hot-deck imputation is similar to the random imputation within-class method when donors are randomly selected with replacement. If the data set to be imputed has no inherent order (i.e., the records in the data file are random), the two procedures are essentially equivalent except for the start-up process. If the data set does have an inherent order, the sequential hot-deck imputation benefits from the positive correlation between donors and recipients. This benefit, however, is unlikely to be substantial.

The advantage of the sequential hot-deck imputation is that all imputations are made from a single pass of the data. However, a problem occurs when the imputation class does not contain an adequate number of donors. An imputation class with too few donors will cause the same donor values to be used repeatedly, creating spikes in univariate distribution of the variables of interest and resulting in a loss of precision in the survey estimates.

Hierarchical Hot-Deck Imputation

Hierarchical hot-deck imputation avoids the disadvantage of sequential hot-deck imputation. This method sorts respondents and nonrespondents into a large number of imputation classes based on a detailed categorization of a large set of auxiliary variables. Nonrespondents are first matched with respondents in the smallest class. If no match is found within that imputation class, classes are then collapsed until a donor is found.

Nearest Neighbor Imputation

Besides sequential and hierarchical hot-deck imputation methods, there are other ways to match a nonrespondent with a donor. For instance, a nonrespondent can be matched to a "nearest" donor, where "nearest" is defined in terms of a statistical distance function based on auxiliary variables. This method is also called *distance function matching* or *nearest neighbor imputation*. The distance function can take many forms. With one auxiliary variable, the distance function can be defined as the minimum absolute differences between the nonrespondent's and donor's values on the auxiliary variable. When multiple auxiliary variables are used, the distance function can be defined as the weighted absolute difference in ranks on auxiliary variables, where the weights represent the importance of the auxiliary variables. A variant of nearest neighbor imputation assigns to the nonrespondent the average value of the neighboring donors.

Assumptions

Hot-deck imputation methods assume that the missing data pattern is missing at random (MAR) within each imputation class; that is, conditional on the auxiliary variables that make up the imputation classes, nonrespondents are no different than respondents. Given this MAR assumption, the selection of auxiliary variables should satisfy two conditions in order for the imputation to reduce nonresponse bias: They have to be correlated with the *y* variable to be imputed, and they have to be correlated with the missing mechanism (i.e., why these values are missing).

Advantages and Disadvantages

Regardless of the specific matching methods, all hot-deck procedures take imputed values from real respondents in the same data file. These procedures have advantages over other imputation methods. For instance, the imputed values are plausible and are within the same range as in the observed data. Hot-deck procedures can reduce bias in univariate statistics such as the mean, and they also have the advantage that they can be carried out as the data are being collected, using everything that is in the data set so far.

There are disadvantages as well. The imputed values, though within the right range, are not necessarily internally consistent for the nonrespondents. Similar to other imputation methods, hot-deck procedures increase the variance of the estimates and may attenuate multivariate relationships. Thus, one must evaluate the hot-deck imputation for any specific data analysis purpose.

Ting Yan

See also Auxiliary Variable; Imputation; Missing Data; Nonresponse; Post-Survey Adjustments

Further Readings

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*, 215–238.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, *16*, 113–131.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Section on Survey Research Methods* (pp. 721–726). Alexandria, VA: American Statistical Association.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*, 1–16.

Household Refusal

The household refusal disposition is used in telephone, in-person, and mail surveys to categorize a case in which contact has been made with a household, but someone in the household has refused either a request by an interviewer to complete an interview (telephone or in-person survey) or a mailed request to complete and return a questionnaire (mail survey). The household refusal typically occurs before a designated respondent is selected. Household refusals are considered eligible cases in calculating response and cooperation rates.

In a telephone survey, a case is coded with the household refusal disposition when an interviewer dials a telephone number, reaches a person, and begins the introductory script, and the person who answers the telephone declines to complete the interview. In calls ending in a household refusal, the person the interviewer spoke with may provide an explanation for the refusal, such as “We don’t do surveys,” “I don’t have time,” “We’re not interested,” or “Please take us off your list.” In other instances, the person contacted may simply hang up. It is important to note that for a case to be coded as a household refusal, the refusal either must occur before the interviewer selects the designated respondent or must be generated by a household member other than the designated respondent. If a refusal was generated by the person known to be the designated respondent, the case should be coded with the respondent refusal disposition, not the household refusal disposition. Past research has shown that the majority of refusals in a telephone survey come from household refusals.

Household refusals in an in-person survey occur when an interviewer contacts a household, a household member answers the door, the interviewer begins the introductory script, and the person declines to proceed with the survey request. As in a telephone survey, cases should be considered household refusals when the refusal occurs before the interviewer selects a designated respondent or when the refusal is provided by a household member other than the designated respondent. A case in an in-person survey should be coded with the respondent refusal disposition—not the household refusal disposition—if a refusal was generated by the person known to be the designated respondent. Common reasons in in-person surveys for household refusals parallel those listed earlier in this entry for telephone surveys.

Cases in a mail survey of specifically named persons are coded with the household refusal disposition when contact has been made with the housing unit in which the sampled person lives and another member of the household declines to have the sampled person complete and return the questionnaire. Because little may be known in a mail survey about who in the household generated the refusal, it can be very difficult, if not impossible, to determine whether a household refusal or respondent refusal disposition is most appropriate, but when in doubt a household refusal should be coded.

Household refusals are considered final dispositions, unless a refusal conversion process is used in the survey. Because refusal rates for all types of surveys have increased significantly in the past decade, many survey organizations review cases ending in household refusals and choose such cases in which the refusal is not extremely strong in nature to be contacted again in order to try to convert the case's disposition to a completed interview.

Matthew Courser

See also Final Dispositions; Refusal Conversion; Respondent Refusal; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

HTML Boxes

Hypertext markup language (HTML) boxes are used in Web-based survey applications and come in all shapes and sizes, but they all allow respondents to Internet surveys to directly input text into an HTML application, for example, an answer to an open-ended survey question. This is extremely useful in survey research for gathering information, such as a respondent's name, email address, and password, but just like other HTML applications, text boxes can be modified to fit one's needs. Fortunately, they all require the same basic parameters, an *input type* and a *form action*. The input type parameter allows one to specify the HTML box one would like to use, whereas the form action defines what will eventually be done with respondent's information.

A one-line box is the most basic HTML box that can be created, as it constrains respondents to using just one input line. These boxes are useful for information responses that are made up of a word or a few words, but they become more cumbersome when used for longer text entries. If one needed to have a larger input space, then a *text area box* would be most appropriate, as the size and shape of this box can be customized allowing it to be used for all sorts of things, ranging from a sentence to much longer, open-ended answers.

In addition to these basic text boxes one could use a *selection box*, which gives the respondent a series of options from which he or she can choose. For online surveys these types of boxes are most commonly used and come in many varieties. For instance, radio buttons allow the users to select only one answer from a provided list; this is useful for survey research, because it prevents respondents from making multiple selections when only one answer is desired by the researcher. *Drop-down selection boxes* can do the same thing, but they look a little different. With drop-down selection boxes, respondents are allowed to select from a dynamic list of items instead of just clicking on the radio button that stands next to the appropriate response. These types of boxes are particularly useful for variables such as income and education, which require a single choice from a list of many options. If one wanted to allow respondents to be able to select multiple items, a *check box* is the easiest way to achieve this end, as they allow users to check as many boxes as they wish. This makes check boxes suitable for measuring constructs such as medical illness histories, which require users to provide several pieces of information for the same question.

Once the preferred input type is chosen, one must then define the form action, which is typically initiated by a "Submit" button that appears at the bottom of the form. Once someone selects the Submit button, a number of things can be done with the form's input, for example, writing the information into a text file and saving it for later use. In addition to a Submit button, often a "Reset" button is added, which allows the user to clear the form's input fields. Even though a Submit button is required to complete the HTML box and a Reset button can be omitted, the Reset button is usually added for user convenience.

Although HTML is a programming language, there are many software editors that are relatively easy to use and achieve the same end. For example, both Microsoft's FrontPage and Mozilla's SeaMonkey provide programmers with a point-and-click interface that allows one to easily create basic and more advanced forms. Additionally, numerous online tutorials are available that provide examples and other resources. Ultimately, HTML boxes are useful for a variety of purposes and always follow the same basic programming model, making them easy to learn and create.

Bryce J. Dietrich

See also Internet Surveys; Open-Ended Question; Radio Buttons

Further Readings

Musciano, C., & Kennedy, B. (2007). *HTML and XHTML: The definitive guide* (6th ed.). Cambridge, MA: O'Reilly Media.

Schonlau, M. (2002). *Conducting research surveys via e-mail and the web*. Santa Monica, CA: RAND.

Sue, V., & Ritter, L. A. (2007). *Conducting online surveys*. Thousand Oaks, CA: Sage.

I

IGNORABLE NONRESPONSE

Researchers who use survey data often assume that nonresponse (either unit or item nonresponse) in the survey is ignorable. That is, data that are gathered from responders to the survey are often used to make inferences about a more general population. This implies that the units with missing or incomplete data are a random subsample of the original sample and do not differ from the population at large in any appreciable (i.e., meaningful and nonignorable) way. By definition, if nonresponse is ignorable for certain variables, then it does not contribute to bias in the estimates of those variables.

Because nonresponse error (bias) is a function of both the nonresponse rate and the difference between respondents and nonrespondents on the statistic of interest, it is possible for high nonresponse rates to yield low nonresponse errors (if the difference between respondents and nonrespondents is quite small). The important question, however, is whether there truly are no meaningful differences between respondents and nonrespondents for the variables of interest. In a major article on this topic, reported in 2006 by Robert M. Groves, no consistent patterns were found between the amount of nonresponse and the amount of nonresponse bias across the myriad surveys that were investigated. That is, in many cases the nonresponse was ignorable and in others it surely was not, and this happened regardless of whether there was a great deal of nonresponse or very little.

The survey response rate is an often-used criterion for evaluating survey data quality. The general and conservative underlying assumption of this is that nonresponse is *not* ignorable. To achieve high response rates, survey organizations must devote a great deal of resources to minimize nonresponse. They might lengthen the field period for data collection, use expensive locating sources to find sample members, use multiple and more expensive modes of contact, and devote additional resources (e.g., through incentives) to convince sample members to cooperate with the survey request. Complex statistical techniques may also be used after data collection to compensate for nonresponse bias. All of these techniques dramatically increase the cost of conducting surveys. In light of this, recent trends of increasing survey nonresponse make the questions of if and when nonresponse is ignorable especially important. If nonresponse does not yield biased estimates, then by implication, it is not advantageous to spend additional resources on minimizing it.

It is difficult to conduct research that evaluates nonresponse error because data for nonresponders to the survey have to be available from some other source. When available, administrative records can be used to evaluate assumptions about nonresponders. However, such studies are rare and expensive to conduct. Other methods used to evaluate nonresponse error include comparing hard-to-reach respondents with easy-to-reach and cooperative respondents or comparing estimates in surveys with identical questionnaires but different response rates.

Though there is relatively sparse evidence that measures nonresponse error in large surveys, nonresponse error in public opinion polls has received some attention in recent years due to the political and media attention focused on such surveys. Public opinion research (especially pre-election polling) usually has a condensed field period that makes a high response rate unattainable. Key variables in these studies include commonly used measures of political and social attitudes and electoral behavior (e.g., party affiliation, ideology, media use, knowledge, engagement in politics, social integration). Most research has found few, or at most minimal (ignorable), differences in the measurement of these variables between surveys conducted in short time spans (approximately 1 week or less) with low response rates (approximately 20% to 30%) and surveys conducted with longer field periods (several months) and higher response rates (approximately 60% to 70%). With respect to sample composition, comparisons between low- and high-response rate surveys often show that both types yield similar estimates on most sociodemographic variables to data from the U.S. Census and other large government surveys. If judged by their accuracy in forecasting elections, many public opinion polls with short field periods and low response appear to be accurate and unbiased. This evidence leaves many researchers fairly confident that nonresponse often may be ignorable for public opinion surveys and that it is unnecessary and inefficient to increase the response rate.

However, these findings are not always consistent. Occasionally, a single but important variable such as party affiliation or political engagement will vary among surveys with different response rates. Also, studies aimed at measuring nonresponse error are still plagued by nonresponse. For example, comparing hard-to-reach respondents with easy-to-reach and cooperative respondents still begs the question of how different the nonrespondents that remain completely unobserved are from the hard-to-reach respondents. Also, analyses that compare low-response rate studies with high-response rate studies implicitly assume that the high-response rate studies do not suffer from nonresponse error. However, even a survey response rate of 70% still leaves 30% of the sample unmeasured.

If characteristics that affect the survey participation decision (e.g., locatability, survey topic, burden, sponsorship, and interviewer behavior, among others) are correlated with variables being measured in the survey, then the survey statistics will change as the response rate increases. For example, evidence from the American Time Use Survey (a large U.S. government survey) finds that busy sample members are no less likely to participate in the survey than others, but people who are weakly integrated into their communities are less likely to respond, primarily because they are less likely to be contacted. If social engagement is an important analytical variable, then nonresponse is not ignorable. Although the evidence from public opinion surveys on ignorable nonresponse may be reassuring, it should not be generalized to surveys focusing on other, nonpolitical topics.

Thus, the issue of when nonresponse is ignorable and when it is not remains a key concern in the field of survey research.

Danna Basson

See also Nonignorable Nonresponse; Nonresponse Error; Nonresponse Rates

Further Readings

Curtin, R. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*, 87–98.

Groves, R. M. (2006). Nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Keeter, S., Miller, C., Kohut, A., Groves, R., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, *69*, 125–148.

Merkle, D., & Edelman, M. (2002). Nonresponse in exit polls: A comparative analysis. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse*. New York: Wiley.

Singer E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 637–645.

IMPUTATION

Imputation, also called ascription, is a statistical process that statisticians, survey researchers, and other scientists use to replace data that are missing from a data set due to item nonresponse. Researchers do imputation to improve the accuracy of their data sets.

Missing data are a common problem with most databases, and there are several approaches for handling this problem. Imputation fills in missing values, and the resultant completed data set is then analyzed as if it were complete. Multiple imputation is a method for reflecting the added uncertainty due to the fact that imputed values are not actual values, and yet still allow the idea of complete-data methods to analyze each data set completed by imputation. In general, multiple imputation can lead to valid inferences from imputed data. Valid inferences are those that satisfy three frequentist criteria:

1. Approximately unbiased estimates of population estimands (e.g., means, correlation coefficients)
2. Interval estimates with at least their nominal coverage (e.g., 95% intervals for a population mean should cover the true population mean at least 95% of the time)
3. Tests of significance that should reject at their nominal level or less frequently when the null hypothesis is true (e.g., a 5% test of a zero population correlation that should reject at most 5% of the time when the population correlation is zero)

Among valid procedures, those that give the shortest intervals or most powerful tests are preferable.

Missing-Data Mechanisms and Ignorability

Missing-data mechanisms were formalized by Donald B. Rubin in the mid-1970s, and subsequent statistical literature distinguishes three cases: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) not missing at random (NMAR). This terminology is consistent with much older terminology in classical experimental design for completely randomized, randomized, and not randomized studies. Letting Y be the N (units) by P (variables) matrix of complete data and R be the N by P matrix of indicator variables for observed and missing values in Y, the missing data mechanism gives the probability of R given Y and possible parameters governing this process, $\xi : p(R|Y, \xi)$.

MCAR

Here, "missingness" does not depend on any data values, missing or observed: $p(R|Y, \xi) = p(R|\xi)$. MCAR can be unrealistically restrictive and can be contradicted by the observed data, for example, when men are observed to have a higher rate of missing data on post-operative blood pressure than are women.

MAR

Missingness, in this case, depends only on observed values, not on any missing values: $p(R|Y, \xi) = p(R|Y_{obs}, \xi)$, where Y_{obs} are observed values in Y, $Y = (Y_{obs}, Y_{mis})$, with Y_{mis} the missing values in Y. Thus, if the value of blood pressure at the end of a clinical trial is more likely to be missing when some previously observed values of blood pressure are high, and given these, the probability of missingness is independent of the missing value of blood pressure at the end of the trial, the missingness mechanism is MAR.

NMAR

If, even given the observed values, missingness still depends on data values that are missing, the missing data are NMAR: $p(R|Y, \xi) \neq p(R|Y_{obs}, \xi)$. This could be the case, for example, if people with higher final blood pressure tend to be more likely to be missing this value than people with lower final blood pressure, even though they have the exact same observed values of race, education, and all previous blood pressure measurements. The richer the data set is in terms of observed variables, the more plausible the MAR assumption becomes.

Ignorability

In addition to defining formally the concepts underlying MCAR, MAR, and NMAR, Rubin defined the concept of *ignorability*. Let the distribution of the complete data, Y, governed by parameter ψ be $p(Y|\psi)$. If (a) the missing data are MAR and (b) ψ and ξ are distinct (which means, in disjoint parameter spaces and, if Bayesian models are used, a priori independent), then valid inferences about ψ can be obtained using a likelihood function that is proportional to $p(Y_{obs}|\psi) = \int p(Y|\psi) dY_{mis}$, and thus, the missing-data mechanism may be "ignored" in likelihood or Bayesian inferences. In many cases, it is reasonable to assume that ψ and ξ are distinct, so that the practical question of whether the missing-data mechanism is ignorable often reduces to a question of whether the missing data are MAR. This argument requires some care, however, when using random parameter models, where ambiguity can exist between parameters and missing data. Also, even when the parameters are not distinct, if the missing data are MAR, then inferences based on the likelihood ignoring the missing-data mechanism are still potentially valid in the sense of satisfying the three frequentist criteria mentioned earlier, but may not be fully efficient. Thus, the MAR condition is typically regarded as more important when considering ignorability.

Single Imputation

Single imputation refers to imputing one value for each missing datum, where the resulting completed data set is analyzed using standard complete-data methods. R. J. A. Little and Rubin offer the following guidelines for creating imputations. They should be

1. Conditional on observed variables
2. Multivariate, to reflect associations among missing variables
3. Randomly drawn from their joint predictive distribution rather than set equal to expectations to ensure that correct variability is reflected

Methods for single imputation typically assume ignorability, and for descriptive simplicity, discussion here is focused on this case.

Unconditional mean imputation, which replaces each missing value with the mean of the observed values of that variable, meets none of the three guidelines for imputation. *Regression imputation* can satisfy the first two guidelines by replacing the missing values for each variable with the values predicted from a regression (e.g., least squares, logistic) of that variable on other variables, but not the third. *Stochastic regression imputation* adds random noise to the value predicted by the regression model and, when done properly, can meet all three guidelines for single imputation.

Hot-deck imputation replaces each missing value with a random draw from a "donor pool" consisting of values of that variable observed on units similar to the unit with the missing value. Donor pools are selected, for example, by choosing units with complete data that have similar observed values to the unit with missing data, for example, by exact matching on their observed values or using a distance measure (metric) on observed variables to define "similar." When the distance is defined as the difference between units on the predicted value of the variable to be imputed, the imputation procedure is termed *predictive mean matching imputation*. Hot-deck imputation, when done properly, can also satisfy all three of the guidelines listed earlier for single imputation.

When single imputations have been created following the three guidelines mentioned earlier, a complete-data analysis of the resulting completed data set can satisfy the first frequentist criterion for valid inferences, but any resulting analyses will nearly always result in estimated standard errors that are too small, confidence intervals that are too narrow, and *p*-values that are too significant, regardless of how the imputations were created, thus failing to satisfy the other two frequentist criteria. Special methods for sampling variance estimation following single imputation have been developed for specific imputation procedures and estimation problems, as described by J. L. Schafer and N. Schenker. However, such techniques need to be customized to the imputation method used and to the analysis methods at hand, and they require the user to have information not typically available in shared data sets. A more broadly applicable, but computationally intensive, approach for sampling variance estimation with singly imputed data is to use a replication technique, such as balanced repeated replication, the jackknife method, or the bootstrap method, with the imputation procedure repeated separately for each replicate. However, such methods assume the first frequentist criterion has been satisfied by the single imputation method, and that the user can conduct all of the repeated imputations.

Multiple imputation, which is discussed in more detail in the next section, is a generally valid approach (i.e., it satisfies the three frequentist criteria) and is broadly applicable and much less computationally intensive for the user than the replication approach. Multiple imputation involves repeating the drawing of single imputation several times, but for its exact validity it requires that the imputations are "proper" or, more generally, "confidence proper." For notational simplicity, ignorability of the missing data mechanism is assumed, even though the ignorability assumption is not necessary for multiple imputation to be appropriate.

A proper imputation is often most easily obtained as a random draw from the "posterior predictive distribution" of the missing data given the observed data, which formally can be written as $p(Y_{mis}|Y_{obs}) = \int p(Y_{mis}, \psi|Y_{obs})d\psi = \int p(Y_{mis}|Y_{obs}, \psi)p(\psi|Y_{obs})d\psi$. This expression effectively gives the distribution of the missing values, Y_{mis}, given the observed values, Y_{obs}, under a model for Y governed by ψ, $p(Y|\psi)$ $p(\psi)$, where $p(\psi)$ is the prior distribution on ψ. The distribution $p(Y_{mis}|Y_{obs})$ is called "posterior" because it is conditional on the observed Y_{obs}, and it is called "predictive" because it predicts the missing Y_{mis}. It is "proper" because it reflects all uncertainty, including in parameter estimation, by taking draws of ψ from its posterior distribution, $p(\psi|Y_{obs})$, before using ψ to impute the missing data, Y_{mis}, from $p(Y_{mis}|Y_{obs}, \psi)$. Rubin has labeled imputation methods that do not account for all sources of variability as "improper." Thus, for example, fixing ψ at a point estimate $\hat{\psi}$, and then drawing an imputation for Y_{mis} from density $p(Y_{mis}|Y_{obs}, \psi = \hat{\psi})$, would constitute an improper imputation procedure.

For simple patterns of missing data, such as with only one variable subject to missingness, the two-step paradigm of drawing ψ from $p(\psi|Y_{obs})$ and then drawing Y_{mis} from $p(Y_{mis}|Y_{obs}, \psi)$ is straightforward to implement. For a simple example, Rubin and Schenker described its use in the context of fully parametric imputation involving logistic regression models. These steps can also incorporate more nonparametric analogues. The simple hot-deck procedure that randomly draws imputations for incomplete cases from matching complete cases is not proper because it ignores the sampling variability due to the fact that the population distribution of complete cases is not known but rather estimated from the complete cases in the sample. Rubin and Schenker described a two-step procedure, termed *approximate Bayesian bootstrap imputation*, which draws a bootstrap sample from the complete cases and then draws imputations randomly from the bootstrap sample.

If the missing data follow a monotone pattern, it is straightforward to draw random samples from $p(Y_{mis}|Y_{obs})$; a pattern is monotone if the rows and columns of the data matrix can be sorted so that an irregular staircase separates Y_{obs} and Y_{mis}. Missing data in clinical trials are often monotone or nearly monotone when data are missing as a result of patient dropout; that is, once a patient drops out, the patient never returns and all data thereafter are missing. Similarly, some longitudinal surveys have monotone or nearly monotone missingness patterns when people who drop out never return. Let Y_0 represent fully observed variables, Y_1 the incompletely observed variable with the fewest missing values, Y_2 the variable with the second fewest missing values, and so on, and assume a monotone pattern of missingness. Proper imputation with a monotone missing data pattern begins by fitting an appropriate model to predict Y_1 from Y_0 and then using this model to impute the missing values in Y_1. For example, first fit a least squares regression of Y_1 on Y_0 using the units with Y_1 observed, then draw the regression parameters of this model from their posterior distribution, and then draw the missing values of Y_1 given these drawn parameters and the observed values of Y_0. Next impute the missing values for Y_2 using Y_0 and the observed and imputed values of Y_1; if Y_2 is dichotomous, use a logistic regression model for Y_2 given (Y_0, Y_1). Continue to impute the next most complete variable until all missing values have been imputed. The collection of imputed values is a proper imputation of the missing data, Y_{mis}, under this model, and the collection of univariate prediction models defines the implied full imputation model, $p(Y_{mis}|Y_{obs})$.

Creating imputations when the missing data pattern is nonmonotone generally involves iteration because the distribution $p(Y_{mis}|Y_{obs})$ is often difficult to draw from directly. However, the *data augmentation* (DA) algorithm, a stochastic version of the expectation-maximization algorithm, is often straightforward to implement. Briefly, DA involves iterating between randomly sampling missing data, given a current draw of the model parameters, and randomly sampling model parameters, given a current draw of the missing data. The draws of Y_{mis} form a Markov chain whose stationary distribution is $p(Y_{mis}|Y_{obs})$. Thus, once the Markov chain has reached effective convergence, a draw of Y_{mis} obtained by DA is effectively a single proper imputation of the missing data from the correct target distribution $p(Y_{mis}|Y_{obs})$, the posterior predictive distribution of Y_{mis}. Many software programs use DA or variants to impute missing values.

An alternative to doing imputation under one specified model is to do imputation under potentially incompatible models, for example, a potentially incompatible Gibbs sampler. These iterative simulation methods run a regression (e.g., least squares, logistic) on each variable having some missing data on all other variables with previously imputed values for these other variables, and then the methods cycle through each variable. Such regression imputation methods have been more extensively developed recently, and they provide very flexible tools for creating imputations and have a relatively long history of application.

Multiple Imputation

Multiple imputation (MI) was introduced by Rubin in 1978. It is an approach that retains the advantages of single imputation while allowing the uncertainty due to the process of imputation to be directly assessed by the analyst using only complete-data software, thereby leading to valid inferences in many situations. MI is a simulation technique that replaces the missing values Y_{mis} with $m > 1$ plausible values, where each single imputation Y_{mis} creates a completed data set, and thus MI creates m "completed" data sets: $Y^{(1)}, \ldots, Y^{(l)}, \ldots, Y^{(m)}$, where $Y^{(l)} = (Y_{obs}, Y_{mis}^{(l)})$. Typically m is fairly small; $m = 5$ is a standard number of imputations to use. Each of the m completed data sets is then analyzed as if there were no missing data, just as with single imputation, and the results of the m analyses are combined using simple rules described shortly. Obtaining proper MIs is no more difficult than obtaining a single proper imputation because the process for obtaining a proper single imputation is simply repeated independently m times. MIs can be created under both ignorable and nonignorable models for missingness, although the use of ignorable models has been the norm.

We now address how to analyze a multiple-imputed data set. Let θ represent the scalar estimand of interest (e.g., the mean of a variable, a relative risk, the intention-to-treat effect), let $\hat{\theta}$ represent the standard complete-data estimator of θ (i.e., the quantity calculated treating all imputed values of Y_{mis} as observed data), and let $\hat{V}(\hat{\theta})$ represent the standard complete-data estimated sampling variance of $\hat{\theta}$. Suppose MI has been used to create m completed data sets. A standard complete-data analysis of $Y^{(l)}$ will create the completed-data statistics, $\hat{\theta}_l$ and $\hat{V}_l = \hat{V}(\hat{\theta})_l$, $l = 1, \ldots, m$, which are combined to produce the final point estimate $\hat{\theta}_{MI} = m^{-1} \sum_{l=1}^{m} \hat{\theta}_l$ and its estimated sampling variance $T = W + (1 + m^{-1})B$, where $W = m^{-1} \sum_{l=1}^{m} \hat{V}_l$ is the "within-imputation" variance, $B = (m-1)^{-1} \sum_{l=1}^{m} (\hat{\theta}_l - \hat{\theta}_{MI})^2$ is the "between-imputation" variance; the factor $(1 + m^{-1})$ reflects the fact that only a finite number of completed-data estimates $\hat{\theta}_l$, $l = 1, \ldots, m$ are averaged together to obtain the final point estimate, and the quantity $\hat{\gamma} = (1 + m^{-1})B/T$ estimates the fraction of information about θ that is missing due to missing data.

Inferences from multiply imputed data are based on $\hat{\theta}_{MI}$, T, and a student's t reference distribution. Thus, for example, interval estimates for θ have the form $\hat{\theta}_{MI} \pm t(1-\alpha/2)\sqrt{T}$, where $t(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the t-distribution. Rubin and Schenker provided the approximate value $v_{RS} = (m-1)\hat{\gamma}^{-2}$ for the degrees of freedom of the t-distribution, under the assumption that with complete data, a normal reference distribution would have been appropriate. J. Barnard and Rubin relaxed the assumption of Rubin and Schenker to allow for a t reference distribution with complete data, and proposed the value $v_{BR} = (v_{RS}^{-1} + \hat{v}_{obs}^{-1})^{-1}$ for the degrees of freedom in the multiple-imputation analysis, where $\hat{v}_{obs} = (1-\hat{\gamma})(v_{com})(v_{com}+1)(v_{com}+3)$, and v_{com} denotes the complete-data degrees of freedom. Rubin and Schenker summarize additional methods for combining vector-valued estimates, significance levels, and likelihood ratio statistics.

A feature of MI that is especially attractive, in the context of data sets that are shared by many users, is that the implicit or explicit model used for imputation, that is, that leads to $p(Y_{mis}|Y_{obs})$, need not be the same as the explicit or implicit model used in subsequent analyses of the completed data. Thus, for example, an organization distributing public-use data can do its best job at imputing for missing data, and then secondary analysts are free to apply a variety of analyses to the multiply imputed data. The formal derivation of procedures for analyzing multiply imputed data, however, is based on the assumption that the imputer's and analyst's models are compatible, in the sense that the imputation model is proper or confidence proper for the analyst's statistics. Formally, as X. L. Meng noted, the imputer's and analyst's models must be "congenial" for the resulting analyses to be fully valid. In order to promote near-congeniality of the imputer's and a variety of user's implicit models, the imputer should include as rich a set of variables in the imputation model as possible. For example, with data from a complex sample survey, variables reflecting features of the sample design, such as sampling weights, or reflecting important domains such as male/female, should be included as well.

It is common to make the ignorability assumption, even when it is not known to be correct, when using MI because (1) it can simplify analyses greatly; (2) the MAR assumption often seems reasonable, especially when there are fully observed covariates available in the analysis to "explain" the reasons for the missingness; (3) MAR cannot be contradicted by the observed data without the incorporation of external assumptions such as exact normality of variables; (4) even when the missing data are NMAR, an analysis based on the assumption of MAR can be helpful in reducing bias by effectively imputing missing data using relationships that are observed; and (5) it is usually not at all easy to specify a correct nonignorable model, and answers can be quite sensitive to its exact form. Therefore, a sensible approach is to use ignorability as a "baseline" assumption and to conduct additional sensitivity analyses using nonignorable models. Rubin has recommended the creation of imputations under multiple models for purposes of sensitivity analysis, in addition to the creation of repeated imputations under a single model for assessments of variability due to missing data under that model.

Many standard statistical software packages now have built-in or add-on functions for creating and analyzing multiply imputed data sets. Routines for creating such data sets include, for example, the S-plus libraries NORM, CAT, MIX, and PAN, for multiply imputing normal, categorical, mixed, and panel data, respectively, which are freely available (see http://www.stat.psu.edu/~jls/misoftwa.html). NORM is also available as a stand-alone version, as is MICE—MI by chained equations (see http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm). In addition, IVEware is very flexible and freely available; it can be called using SAS or can be run as a stand-alone version (http://www.isr.umich.edu/src/smp/ive/). SAS now has procedures PROC MI and PROC MIANALYZE making the analysis of multiply imputed data sets easy. Other software packages have been developed specifically for creating multiply imputed data sets; for example, the commercially available SOLAS (http://www.statsol.ie/solas/solas.htm) is most appropriate for data sets with a monotone or nearly monotone pattern of missing data. Additionally, STATA provides MI routines based on the chained equation approach and supports analyses of multiply imputed data sets. For more information, see www.multiple-imputation.com.

Susanne Rässler, Donald B. Rubin, and Elizabeth R. Zell

See also Balanced Repeated Replication (BRR); Bootstrapping; Hot-Deck Imputation; Jackknife Variance Estimation; Missing Data; Multiple Imputation; SAS; Stata

Further Readings

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 948–955.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). *Survey nonresponse*. New York: Wiley.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Madow, W. G., Olkin, I., & Rubin, D. B. (1983). *Incomplete data in sample surveys*, *Vol. 2*: *Theory and bibliographies*. New York: Academic Press.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, *9*, 538–573.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. New York: Springer.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–590.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*, 538–543.

Rubin, D. B. (1978). Multiple imputation in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods* (pp. 20–40). Alexandria, VA: American Statistical Association.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (2nd ed.). New York: Wiley.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*, 366–374.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, *95*, 144–154.

Inbound Calling

Telephone survey calls involving call centers are classified as *inbound* or *outbound*, depending on whether the call is being received by the call center (inbound) or initiated in the call center (outbound).

Inbound calling in the survey research context usually arises from one of the following situations:

- As support for an existing mail, Web, computer-assisted telephone interview (CATI), or in-person survey. In a pre-survey letter or email to the respondent, or on the actual instrument in the case of a self-completion survey, a phone number is provided for respondents to call if they have any questions about the survey or if they want to schedule a particular time for an interviewer to call. Depending on the survey design, respondents might also be given the option to complete the survey via interview during the inbound call.
- As an additional data collection mode in a multi-mode survey. Longitudinal surveys, where a reliable mail or email address exists for communication, often offer respondents an incentive to phone in and complete the interview at their own convenience before the main field phase. Web and mail surveys often contain a phone number for respondents to call if they would prefer an interview to a self-completion mode.
- As the only data collection option, as might be appropriate for surveys of

- People who visit a particular location or use a particular service, when information is sought about their experience or transaction
- Rare populations for which no sampling frame exists and screening is not feasible
- Surveys on a particularly sensitive topic for which respondents are more likely to give accurate information only if they can do so anonymously (i.e., without the interviewer knowing who is calling). Because in these types of situations the probability of inclusion is rarely calculable, minimum incidence rates, as opposed to population estimates, are usually sought. Distributing the phone number for respondents to call in such cases is typically done by pamphlet, advertising in the press, or via related-interest Web sites and chat rooms.

The technology used for inbound calls can be as simple as a single phone on a single interviewer's desk or something as complex as that used in the large commercial call centers that support commercial banks and credit card companies, with queuing, interactive voice response (IVR), and automatic call distribution systems. Some of this advanced functionality, however, is of limited use in the survey context, because survey respondents have little motivation to wait in queues, and many surveys are longer than is practical for data collection to be done entirely via IVR.

The telephone number provided is usually toll-free (where the receiver pays for the call) to encourage more respondents to phone in. Usually different toll-free numbers are used for different surveys, to allow for more customization in the greeting. Support for multiple languages is done by publishing a different toll-free number for each language or else using a simple IVR system (such as "press 1 for English, 2 for Spanish").

Staffing is determined by both technology and service expectations. For example, more staff are needed if the phone needs to be answered by a live operator within 60 seconds during the advertised hours, and fewer staff are needed for a survey where it is acceptable for the majority of inbound calls to be routed to a voicemail system. Getting the right balance of staff is critical, as too many will result in unproductive use of interviewer time, but too few will lead to unanswered or queued calls, which is likely to irritate the respondent, create more refusals, and thereby reduce response rates.

Blended environments—where interviewers making outbound calls can also receive inbound calls—can assist staffing adequately for inbound callers, but the technology required to make blended environments effective can be complicated.

Record-keeping practices depend on the survey design and the technology available. It is advantageous if the telephony system used can capture caller-ID information, so that even if the call goes to voicemail and the respondent declines to leave a message, some record is still available that will allow, at the very least, for the inbound call to be added to the contact history for that case.

Ideally the inbound system will also be tied to the outbound survey's sample management system, so that when a respondent calls in, the relevant sample case will be accessed and thus prevent some other interviewer from making an outbound call to the same number. In addition, the history of the case can be updated to assist any interviewer who might make a subsequent call to that number.

Jenny Kelly

See also Interactive Voice Response (IVR); Outbound Calling; Sample Management; Telephone Surveys

Further Readings

Koole, G., & Mandelbaum, A. (2002). Queuing models of call centers: An introduction. *Annals of Operations Research*, *113*, 41–59.

INCENTIVES

Providing incentives, such as cash, to potential survey respondents is an effective way to increase response rates and thereby possibly reduce the potential for nonresponse bias. Incentives work best when combined with a multiple contact recruitment approach, but incentives demonstrate their effectiveness in improving response rate even at the time of first contact. For interviewer-mediated surveys, the judicious use of incentives can reduce the number of contacts required to complete an interview. Because incentives increase early responses, incentives have demonstrated their time-efficiency and cost-effectiveness by reducing the labor and postage costs of additional contacts.

Theories of Incentives

Several theories are used to explain why incentives work. The most common explanations rely on social exchange theory, but other theories include economic exchange, cognitive dissonance, and leverage-saliency theory.

Social Exchange Theory

Social exchange theory proposes that when people consider participating in a survey, they evaluate a variety of nonmonetary costs and rewards, such as the desire to help or social validation. A pre-paid incentive, whether cash or noncash, is thought to be a goodwill gesture that puts the survey, the researcher, and the sponsor in a positive light, encouraging compliance with a request to complete a survey. In addition, this gesture may establish trust between the researcher and the respondent, further encouraging compliance. This theory assumes the creation of a sense of obligation or an expectation of reciprocity, which may be facilitated by the trust established between the researcher and respondent. The incentive produces a sense of obligation that a favor needs to be exchanged, regardless of the magnitude of the favor.

This theory encourages the presentation of an incentive as a token of appreciation. Don Dillman, an early proponent of this theory, has cautioned against matching the monetary incentives to the level of effort required to complete and return the survey. This matching could trigger thinking about the relationship as an economic exchange, making it easier for people to refuse. Instead, social exchange theory depends on the perception of the incentive as a token of appreciation or symbol of trust.

Economic Exchange Theory

Economic exchange theory proposes that incentives be used to compensate people for the costs (burden) associated with survey participation, that is, pay respondents for their time and effort. This theory depends on people exercising rational choice in deciding whether to participate in a survey. People weigh the costs of participating (time and effort) against the benefits of participating, such as receiving an incentive. From this perspective, the greater the burden is—that is, the longer and more demanding the survey—the greater the incentive should be.

Cognitive Dissonance

Cognitive dissonance, as explained by social psychologist Leon Festinger and his colleagues in the 1950s, proposes that if people are given a noncontingent (pre-paid) incentive with a survey request, they will be motivated to comply with the request to avoid or to reduce an uneasy feeling that arises from accepting a reward without having done anything to deserve it. When presented with a pre-paid incentive to complete a survey, respondents are faced with several choices. They could keep the incentive and return the completed survey (as requested), or they could return the incentive with the uncompleted survey. Both choices avoid a state of dissonance. On the other hand, if the respondent accepts the incentive but does not return the survey, in theory, they experience dissonance because they did not comply with a request but accepted the reward of the request. This experience of dissonance will be unsettling and may lead to several courses of action at a later date, including returning the incentive or returning the completed survey. Survey researchers hope that most respondents will be inclined to complete and return the survey, because the other courses of action produce unsettled feelings or require as much effort as compliance with little reward—that is, returning the uncompleted survey and incentive.

Leverage-Saliency Theory

Leverage-saliency theory, first advanced in 2000 by Robert Groves and his colleagues, presumes that people are influenced by a variety of factors, which differ in importance (salience), in deciding whether to participate in a survey. Some people are motivated, for example, by survey topic or sponsor, while others are motivated by civic duty or cash incentives. These different factors, if they are known before survey recruitment, could be emphasized during survey recruitment to improve response rates. Groves and colleagues use the analogy of a scale to illustrate leverage-saliency theory. The scale can be tipped in favor of survey participation, if the survey request is tailored to the respondent's key decision-making factors. For some respondents, an incentive will be highly salient and thus serve as an important leveraging factor in determining whether or not they will participate with a survey request.

Noncontingent Incentives Versus Contingent Incentives

Research has found consistently that incentives are most effective when provided prior to the completion of the survey task, that is, when given as pre-paid or noncontingent incentives. Noncontingent incentives are much more effective than incentives of the same size that are provided upon completion of the survey (promised or contingent incentives). In addition, some studies have found that noncontingent incentives produce higher response rates than promised incentives that are of much greater value.

The greater effectiveness of noncontingent incentives, over contingent incentives, offers support for both social exchange theory and cognitive dissonance theory of incentives. The pre-paid incentive may produce a sense of trust or invoke a sense of reciprocity, which explains survey participation. Alternatively (or in addition), potential respondents may be motivated by a desire to avoid a state of dissonance that may arise from keeping an incentive without completing the request.

A meta-analysis conducted by Allan Church suggested that promised incentives do not have any significant effect on response rates. However, the mode of the survey may interact with the effectiveness of contingent versus noncontingent incentives. Another meta-analysis of interviewer-mediated surveys by Eleanor Singer suggested that promised incentives produce a statistically significant increase in response rates and, in some cases, may be as effective as pre-paid incentives in recruiting respondents.

Contingent incentives include those incentives that are awarded as part of a prize drawing or lottery. Like other contingent incentives, prize draw awards generally are less effective than pre-paid incentives, even in conditions where the prize draw has a considerably greater potential payoff value. Research on the effectiveness of prize draws is split into two camps: that they produce slightly higher response rates or they have no effects on improved response rates. Prize draws are attractive, because they are much less costly to administer, although they have other administrative headaches.

Cash Incentives Versus Noncash Incentives

Cash awards are generally more effective than noncash awards (e.g., checks, gift cards, gift certificates, or small tokens such as pens), even if the value of the noncash award is greater than the cash award. Some studies find the greater the cash incentive is, the higher the response rate will be, but other studies find that the relationship is not strictly linear. These conflicting findings suggest that there are linear effects with lower amounts of incentives (where each additional dollar produces a higher response rate) and a flattening effect with higher amounts of incentives (additional dollars produce little change in response rate, e.g., going from $20 to $25). The diminishing return between the size of the incentive and response rate casts doubt on economic exchange theory of incentives as the sole or primary explanation of how incentives work.

While noncash incentives are less effective than monetary awards, some noncontingent, noncash awards produce significant increases in response rates. Pens and charity donations appear to be the favored options for noncash incentives among those that have been reported in the research literature. Pens vary in effectiveness in increasing response rate, but the research on the role of charity contributions is more conclusive. Charity donations on behalf of the respondents do not produce a significant increase in response rates.

Data Quality

Initially, survey researchers worried that incentives would reduce data quality by reducing intrinsic motivation to complete a survey, leading to less thoughtful or less complete responses. Subsequent research has largely appeased this concern. Studies find that incentives improve or have no effect on data quality. When comparing item nonresponse to closed-ended questions, research tends to be equally split between finding no differences and finding fewer item nonresponses in the incentive groups, compared to no-incentive groups. For example, the Singer meta-analysis of interviewer-mediated studies found that about half of the studies had no differences in data quality and half of the studies found improved data quality, when respondents received an incentive. When examining item nonresponse and length of response to open-ended questions, research suggests that incentives lead to higher-quality responses, as indicated by more distinct ideas and longer responses.

Response Bias

Survey researchers worry that incentives might alter survey responses, producing response bias. The possibility of response bias has been evaluated by comparing responses from those respondents who receive an incentive to those respondents who do not receive an incentive. (These investigations do not rule out the nonresponse bias that may exist between respondents and nonrespondents.) In general, differences in responses between incentive groups and no-incentive groups are not statistically significant. However, a few studies have found that incentives may produce slightly more favorable evaluations or positive ratings from incentive groups compared to no-incentive groups, but the small differences may not be substantively meaningful.

Sample Bias

In general, studies have found few differences in the demographic composition of their samples, when comparing no-incentive groups and those groups who received incentives of various types and sizes. Thus, incentives appear to be effective in recruiting respondents from a broad range of demographic groups. Unfortunately, these findings also suggest that incentives may not be effective in recruiting respondents who are especially hard to reach or otherwise unlikely to participate in surveys (e.g., young adults), unless differential amounts of incentives are given to such subgroups. For example, Singer and colleagues reported that only one third of the studies in their meta-analysis provided support for the role of incentives in improving the representation of groups that typically are underrepresented (low-income and nonwhite) in surveys, but these studies did not provide differential incentives to these groups. Most of the interviewer-mediated studies affirm that incentives do not change the demographic composition between incentive and no-incentive groups, unless differential incentives are used. Nor is there reliable and consistent evidence that incentives reduce nonresponse bias.

Interaction Effects of Incentives in Panel Data and Mixed-Mode Studies

The role of incentives in panel or multi-wave studies has increasingly become the subject of investigation. The effectiveness of incentives may differ by the outcome of prior contact with the respondent. While Singer and her colleagues found that incentives work for panel respondents, fresh respondents, and prior refusing respondents, Norm Trussell and Paul Lavrakas found that the improvement in response rate is proportionally much greater for people who refused during a prior contact and later received a larger incentive, than for those with whom the researchers had no contact in the prior stage of the research or those who readily agreed to cooperate in the subsequent research without receiving an incentive. John Brehm found that incentives work better when a respondent refuses because of the survey characteristics (e.g., survey is too long), but incentives may be counterproductive if the respondent refuses because they doubt their ability to do the survey (they are illiterate, deaf, or ill). Using the findings from previous interactions with a respondent to tailor the recruitment strategy, including the targeted use of incentives, is the basic tenet of leverage-saliency theory.

Shelley Boulianne

See also Bias; Contingent Incentives; Economic Exchange Theory; Leverage-Saliency Theory; Missing Data; Noncontingent Incentives; Nonresponse Error; Response Rates; Social Exchange Theory; Tailoring

Further Readings

Brehm, J. (1994). Stubbing our toes for a foot in the door? Prior contact, incentives and survey response. *International Journal of Public Opinion Research*, *6*(1), 45–63.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, *57*(1), 62–79.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailor designed method* (2nd ed.). Hoboken, NJ: Wiley.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, *64*, 299–308.

Hopkins, K. D., & Gullickson, A. R. (1992). Response rates in survey research: A meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education*, *61*(1), 52–62.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., & McGonagle, K. A. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, *15*(2), 217–230.

Trussell, N., & Lavrakas, P. J. (2004). The influence of incremental increases in token cash incentives on mail survey response: Is there an optimal amount? *Public Opinion Quarterly*, *68*(3), 349–367.

Warriner, K., Goyder, J., Gjertsen, H., Hohner, P., & McSpurren, K. (1996). Charities, no; lotteries, no; cash, yes: Main effects and interactions in a Canadian incentives experiment. *Public Opinion Quarterly*, *60*(4), 542–562.

Independent Variable

In survey research, an independent variable is thought to influence, or at least be correlated with, another variable: the dependent variable. For example, researchers hypothesize that childhood exposure to violent television can lead to violent behavior in adulthood. In such a study, exposure to violent television programming as a child is an independent variable and violent behavior in adulthood is the dependent variable.

An independent variable is commonly denoted by an x and a dependent variable by y, with the implication that "x causes y" or, in the case of noncausal covariation, "x is related to y."

Determining whether one variable influences another is of central importance in many surveys and studies, as making this determination helps researchers accept or reject hypotheses and thereby build social science knowledge. Relationships between variables help researchers to describe social phenomena.

In experimental studies, with random assignment of respondents to experimental conditions, a researcher can choose which variables are independent, because these are the variables controlled by the researcher. In population studies, patterns in data help researchers determine which variables are independent.

More than one independent variable may influence a dependent variable. Quantitative tools and approaches can assist researchers in accepting or rejecting their hypotheses about the relationships among independent variables and a dependent variable. In some analyses, researchers will "control for" the influence of certain independent variables in order to determine the strength of the relationship for other independent variables.

Using the example of childhood exposure to television violence again, another independent variable in the study could be parental control over television viewing. Yet another independent variable could be level of physical violence in the home. The complexity of a hypothesized causal model such as this increases with the number of independent variables and interaction effects among independent variables.

Heather H. Boyd

See also Dependent Variable; Experimental Design; Interaction Effect; Noncausal Covariation; Random Assignment

Further Readings

Cohen, J. (Ed.). (2003). *Applied multiple regression/correlation for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.

StatSoft, Inc. (2005). *Electronic statistics textbook*. Retrieved February 26, 2007, from http://www.statsoft.com/textbook/stathome.html

Ineligible

The ineligible disposition is used in all surveys, regardless of the mode (telephone, in-person, mail, or Internet). Ineligible cases are cases that were included in the sampling frame but fail to meet one or more of the criteria for being included in the survey. Cases coded as ineligible do not count in computing survey response rates.

One of the most common reasons a case would be coded as ineligible is when a survey uses screening criteria to determine whether the respondent, household, or organization contacted as part of a survey is eligible to complete the survey. For example, a survey may require respondents or households to be located within a specific geographic area, such as a specific county, town or village, or neighborhood. A case would be considered ineligible if it were discovered that the respondent or household was located outside of the geographic boundaries of the survey population. In most instances, if it were discovered during the screening process that the sampled respondent had moved out of the geographic boundaries of the survey during the field period, that case also would be considered ineligible.

An additional example of how screening criteria may result in a case being considered ineligible occurs in surveys of the general population. These surveys use residential status as screening criteria, and as a result, all cases that result in contact with a nonresidential unit, such as businesses, schools, or governmental organizations, would be considered ineligible. In in-person surveys, this often is discovered when an interviewer visits a sampled address and discovers that it is not a residence. In telephone surveys, this would be discovered when interviewers make telephone calls to businesses; fax or data lines; nonworking, changed, and disconnected telephone numbers and numbers that reach pagers. In landline telephone surveys, numbers that reach cell phones would be treated as ineligible. Also, in a telephone survey an answering machine message might allow an interviewer to determine if the number is ineligible.

Some surveys use screening at the respondent level to determine eligibility. For example, a survey may seek to collect data from respondents with a specific set of characteristics (demographics, occupation, tenure in job, etc.). Cases in which the individual respondent discloses during the screening process that he or she does not have the characteristics sought by the survey would be considered ineligible. Finally, if a telephone, in-person, mail, or Internet survey uses quotas, cases contacted for which quotas have already been filled are considered ineligible.

A number of other reasons that a case may be categorized with the ineligible disposition are specific to each survey mode. In telephone surveys, the ineligible disposition may be used when the number has technical difficulties and no one can be reached on it or when a business number is forwarded to a residence. In an in-person survey, the ineligible disposition may be used for cases in which interviewers discover that the sampled address is a housing unit that is vacant during the entire field period of a survey, and rarely, for a housing unit that has no eligible respondent (such as cases in which all residents are under 18 years of age). In mail and Internet surveys, the ineligible disposition may be used if the same respondent or addressee is sampled more than once. These duplicate mailings usually are treated as ineligible if the error is not caught until after the questionnaires have been mailed out (in mail surveys) or until after the email invitation is sent out (in Internet surveys).

Matthew Courser

See also Final Dispositions; Out of Sample; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

INFERENCE

Inference is a process whereby a conclusion is drawn without complete certainty, but with some degree of probability relative to the evidence on which it is based.

Survey data may be used for description or for analysis. Descriptive uses include making estimates of population totals, averages, and proportions. Analytical uses include testing hypotheses about the population. The analytical uses involve making *statistical inferences*. For example, a descriptive use of survey data would be to supply an estimate of the number of male and female engineers. An analytical use would be to infer (based on valid statistical procedures) that there are significantly fewer female engineers than male engineers. Another descriptive use of survey data would be to supply the average salary of teachers. An inferential use would be to conclude that, even after controlling for education and experience, teachers of one racial-ethnic category tend to have a higher average salary than those in another racial-ethnic category.

Design-Based and Model-Based Inferences

There are two approaches to making inferences from survey data. First, in the *design-based approach*, inferences are made by looking at how statistics vary as samples are repeatedly drawn using the same sampling procedures as were employed in the actual sampling.

Second, in the *model-based approach*, inferences are made by looking at how statistics vary as the population, as described by a probability model, is allowed to vary without changing the sample. The model-based approach is also called the *prediction approach* because the model is used to predict the population units not in the sample. It is called the *superpopulation approach* as well because the population can be regarded as selected from a still larger population according to the probability model.

Inference procedures (e.g., hypothesis testing or estimating confidence intervals) can be carried out under either the design-based or the model-based approach. The design-based approach is more traditional in survey sampling. The model-based approach, on the other hand, is more consistent with statistical approaches used outside of survey sampling.

Confidence Intervals

Confidence intervals allow one to infer with a high degree of confidence that a quantity being estimated lies within an interval computed by a specified procedure.

The precise meaning of "confidence" depends on whether one is adopting the design-based or model-based approach. Clearly a confidence interval is more informative than a numerical estimate of a population quantity (called a *point estimate*) in that the confidence interval conveys information about how precise the point estimate is.

Hypothesis Testing

The purpose of hypothesis testing is to ascertain whether an observed difference in the sample is statistically significant or whether it can instead be adequately explained by chance alone. Hypothesis tests are designed so that, if there is in fact no difference, the probability of (erroneously) rejecting the hypothesis that there is no difference (i.e., the null hypothesis) is kept to a specified low level; often this probability, called the Type I error, is set to .05. A well-designed hypothesis test will also minimize the other potential error on inference, namely, not rejecting the hypothesis of no difference when a difference actually exists (i.e., Type II error). In survey sampling, it is often the case that two sample averages are independent and approximately normally distributed so the hypothesis that their difference is zero can be tested using properties of the normal distribution (this is called a *t-test*).

There is a close relationship between confidence intervals and hypothesis testing in that a hypothesis of no difference is rejected if and only if the confidence interval *for the difference* does not include zero. If one has confidence intervals for each of two independent averages and the confidence intervals do not overlap, one may reject the hypothesis of no difference between the two averages. But if the two confidence intervals do overlap, it is still possible that the hypothesis of no difference in the sample averages can be rejected.

Michael P. Cohen

See also Confidence Interval; Design-Based Estimation; Model-Based Estimation; Null Hypothesis; Point Estimate; Regression Analysis; Superpopulation; *t*-Test; Type I Error; Type II Error

Further Readings

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. Chichester, UK: Wiley.

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. New York: Wiley.

Skinner, C. J., Holt, D., & Smith, T. F. M. (1989). *Analysis of complex surveys*. Chichester, UK: Wiley.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling: A prediction approach*. New York: Wiley.

INFORMANT

An informant in a survey is someone asked to provide information about another person, persons, or an organization, for example, when a parent is interviewed about a young child who could not answer the survey questions. Informants (also known as *proxies*) tend to be used in surveys when the target respondent is unable to respond or when it is not feasible to collect responses from all members of a group under study. As the use of informants to collect quantitative data has become integral to survey research, due to the cost-effectiveness of the approach, so has the study of the effects of using informants on the data that are collected. The substitution of informants limits the types of data that can be collected with accuracy, and proper selection methods should be used to minimize the resulting response bias that has been noted to occur occasionally when data collected from informants are compared to self-reported data.

There are several types of surveys in which the use of informants is the most efficient means of collecting responses. Informants are frequently used in surveys when members of the population under study are unable to provide responses because a physical or cognitive impairment prevents them from responding. Because informants are the only way to collect information about these target populations, informants are used despite the fact that several methodological experiments have shown that using informants produces response bias. Specifically, in surveys asking about disability, informants tend to overreport more obvious types of disability (such as difficulty with activities of daily living) and to underreport less obvious types (such as mental health problems).

Another survey situation in which the informant method of data collection has been used is when it is not economically feasible to interview all individual respondents, such as all members of a household or an organization. Studies using this method generally develop selection rules that ensure that the selected informant is likely to be able to provide accurate responses about others in the household or organization. The selection rules used by the survey are applied (e.g., randomly selecting one of the adults in the household), and the selected informant reports on the behaviors of other members of the household or organization. Further, surveys asking questions on sensitive topics have used informants to reduce bias from self-reported responses due to social desirability. These include surveys collecting information about sensitive behaviors such as alcohol, tobacco, and drug use, which tend to be underreported by respondents. For example, the results of some experiments suggest that the informant method yields estimates of alcohol use that are closer to actual alcohol sales figures than self-reported data.

The ability of selected informants to respond accurately to surveys depends on how observable the survey subjects are by the informant and on the informant's ability to recall events. The survey topics asked of informants must take these issues into account, as informants are best able to provide accurate information about others when the informant has a high degree of knowledge about those he or she is answering questions about. For this reason, topic areas in surveys using informants that include factual or "hard" measures are preferable to "soft" measures requiring subjective evaluation by those providing responses. For example, in a survey administered to principals of an organization, demographic questions answered by an informant on "hard" topics such as age, prior experience, or race will likely produce lower item nonresponse and be more reliable than topics less likely to be observable to informants, such as net worth or total household income.

David DesRoches

See also Respondent-Related Error; Response Bias

Further Readings

Mathiowetz, N. A., & Groves R. M. (1987). The effects of respondent rules on health survey reports. *Vital Health Statistics*, *2*(106), 26–32.

Singer, E. (1973). Agreement between "inaccessible" respondents and informants. *Public Opinion Quarterly*, *36*(4), 603–611.

Todorov, A., & Kirchner, C. (2000). Bias in proxies' reports of disability: Data from the National Health Interview Survey on Disability. *American Journal of Public Health*, *90*(8), 1248–1253.

Informed Consent

As outlined in the *Belmont Report* of 1979, the core elements underlying the ethical treatment of research participants are autonomy (respect for persons), beneficence, and justice. Providing adequate information and obtaining active consent for research participation are central to autonomy and respect for persons. Acknowledging the importance of autonomy requires that every potential research participant must be afforded adequate time and opportunity to make his or her own informed and voluntary decision about whether or not he or she wishes to participate in a research study. This requires the provision of adequate information about the study and, in theory, also requires that no pressure be exerted to participate. The principle of autonomy also requires that special protections be given to potentially vulnerable populations such as minors, the mentally ill, or prisoners. Individuals in these groups may be in a position of increased potential for coercion (e.g., prisoners) or may be less capable of understanding information that would enable them to make an informed decision about study participation.

Informed consent includes the process by which research participants gain an understanding of the procedures, risks, and benefits that may be associated with their taking part in a study. In virtually all surveys, the key elements of "voluntary" and "informed" consent can be provided in a concise way at the beginning of a telephone or face-to-face interview, in a cover letter for a self-administered survey, or in the introductory screen of a Web or other electronic survey. This is true regardless of level of risk and is consistent with the contemporary view of consent as an ongoing process rather than a paper document. The main elements of consent include the following:

- An explanation of the purpose(s) of the study
- An indication of the approximate amount of time it will take to complete the study
- A description of what the respondents will be asked to do
- A description of any foreseeable risks or discomforts, if any
- A description of any direct benefits to the respondents or others
- A statement describing the extent to which responses will be confidential
- A statement of the voluntary nature of participation

Contact information should also be available for questions about the research and about respondent rights. This information can be provided upon request for telephone surveys and should be included in the written introductory information in face-to-face, self-administered, and electronic modes.

Consent can be obtained from adult respondents who can understand the benefits and risks of the survey. Except in the special cases where parental permission itself could pose risks (e.g., studies of child abuse), parental permission must be obtained prior to administration of a survey to a minor, and assent (agreement to participate) should be obtained from the child or other nonadult. The age of majority is typically 18 but varies slightly in the United States from state to state.

Special challenges exist for studies being conducted by or for someone with authority or special power over potential respondents (e.g., teachers, supervisors, employers, physicians). In these cases, it is particularly important that the respondent recruitment procedures evidence no coercion, either explicit or implicit. Researchers must make a clear distinction between research questions and issues arising out of the authority relationship. To avoid the perception of undue influence or coercion, persons with authority over potential respondents should not recruit participants themselves if they will have knowledge about who did and did not participate or will have access to individual responses.

In most surveys, respondents indicate their consent by providing oral agreement at the beginning of the interview, by answering questions as they are asked or that appear on a paper or electronic questionnaire, or both. Thus, people may consent to all of an interview, to part of it, or to none of it, depending on how they respond to requests from the interviewer.

Federal regulations (C.F.R. 46.117c) on human subjects protections in the United States recognize that written or signed consent forms are not necessary or desirable in every research setting. The regulations provide that, while written consent is the norm in much research involving humans, institutional review boards may waive requirements for signed consent if they find that the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context. Most surveys pose risks no greater than those experienced in everyday life. Further, in many contexts, written consent forms may increase risk (e.g., research on illegal behavior, health, illegal immigrants) and may reduce cooperation unnecessarily. Telephone surveys utilizing random-digit dialing and Web survey invitations sent via email cannot incorporate signed consent in the protocol prior to the initial contact because respondents' names and street addresses are unknown to the researcher. Thus, a waiver of documentation of consent is typically the most desirable approach for most survey protocols—especially those utilizing telephone and electronic modes.

In the typical survey that presents minimal risk, lengthy and detailed information about the objectives of the survey and the questions to be asked may increase respondent burden and bias responses without safeguarding respondent rights. In these surveys, the usual practice of a short introduction—including the purpose of the study; the approximate amount of time it will take; the sponsor, responsible survey organization, or both; and the general topics to be covered—is typically deemed sufficient. This statement should also include information about the confidentiality of the responses.

More detailed information is required when survey participation may pose substantial risk. In general, respondents should be informed that the content includes sensitive topics or questions about any illegal behaviors, but they should not be told so much as to bias their answers (e.g., they should not be informed of the study hypothesis). This is consistent with much other social science research performed in laboratory settings where explanations of the hypotheses at the outset would render the study invalid, although such studies may require special debriefing of the respondents after the survey is completed. It is also important to include a statement indicating that respondents can skip questions that cause them discomfort and questions they do not want to answer.

The introductory statement and the reminders about the voluntary nature of response help ensure respondent autonomy (respect for persons) without affecting substantive responses. If appropriate, at the end of the interview, respondents can be debriefed to see if any of the matters covered were upsetting, to give further information on study purposes, or to answer respondent questions.

Mary E. Losch

See also Debriefing; Ethical Principles; Institutional Review Board (IRB); Protection of Human Subjects; Voluntary Participation

Further Readings

Citro, C., Ilgen, D., & Marrett, C. (Eds.). (2003). *Protecting participants and facilitating social and behavioral sciences research*. Washington, DC: National Academies Press. Retrieved May 2003 from http://www.nap.edu/books/0309088526/html
Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart A), 101–124.
Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart B), 201–211.
Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart C), 301–306.
Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart D), 401–409.
Singer, E. (2003). Exploring the meaning of consent: Participation in research and beliefs about risks and benefits. *Journal of Official Statistics*, *19*, 273–285.

Institute for Social Research (ISR)

The Institute for Social Research (ISR), originally established in 1949 at the University of Michigan as the Survey Research Center under the direction of Rensis Likert, is the nation's longest standing interdisciplinary research laboratory in the social sciences. Support from the sociology and psychology departments was an important factor in the decision to establish the ISR at the University of Michigan. The center was kept administratively separate from other schools and colleges, and its executive committee was made up of members from a variety of relevant disciplines; this separation was designed to promote its interdisciplinary nature. By the early 1950s the staff had grown from 12 to nearly 200. Tenure privileges were not granted to the senior staff members by the University of Michigan Regents until 1968. Its research staff exceeds 225, representing the disciplines of psychology, sociology, economics, political science, and others. Nearly 500 permanent staff and additional contingent interviewers numbering 1,300 support the research scientists. The ISR is primarily self-supporting through grants and external contracts, largely from the federal government. ISR study findings have been used to develop policy and practice on various social issues ranging from health and drug use to racial prejudice, welfare, and retirement. The ISR is best understood through its research centers, each autonomous in its research and administration.

The Survey Research Center is the largest of the ISR centers and is a national and international leader in social science research. The four initial programs—economic behavior, political behavior, organizational behavior, and survey methodology—have grown to include active research programs such as family and demography, life course development, quantitative methodology, social environment and health issues, socioenvironmental studies, social indicators, urban and environmental studies, and youth and social issues.

The Center for Political Studies research themes are political values and participation, media and politics, race and politics, and international peace and security. Designated as a national resource by the National Science Foundation, the Center for Political Studies was originated by the development of the National Election Studies, first conducted in 1948.

The Inter-University Consortium for Political and Social Research, created in 1962, has over 600 member institutions worldwide and is the largest archive of quantitative social science data in the United States. The archive includes some 500,000 files for research and instruction. These files span political science, history, public health, sociology, demography, criminal justice, international relations, economics, gerontology, and education.

The Research Center for Group Dynamics focuses on the individual and has particular interest in social cognition and group processes. The center was founded in 1945 at the Massachusetts Institute of Technology and moved under the ISR umbrella in 1948. Studies in natural settings probe decision making, prejudice, emotion, and judgment processes, and research has practical applications to understanding group behavior and social influence that affect social concerns such as racism, crime, and delinquency.

The Population Studies Center was established in 1961 within the Department of Sociology at the University of Michigan and has been closely related to the Department of Economics since 1966. Drawing faculty from the departments of anthropology, biostatistics, economics, geography, history, natural resources, political science, psychology, public health public policy, social work, sociology, and statistics has allowed the Population Studies Center to become increasingly interdisciplinary. The center's strengths in demographic research include aging; family formation, fertility, and children; health, disability, and mortality; human capital, labor, and wealth; methodology; population dynamics; and regional studies. In 1998, the center made an institutional move from the College of Literature, Science, and Arts to ISR, becoming the institute's fourth center.

Training future generations of empirical social scientists is the long-term commitment of the ISR. The ISR has offered the Summer Institute in Survey Research Techniques for more than 60 years; designed to meet the needs of professionals, the Summer Institute teaches practice, theory, implementation, and analysis of surveys. For more than 45 years, the Inter-University Consortium for Political and Social Research (ICPSR) Summer Program has provided studies in basic methodological and technical training, data analysis, research design, social methodology, and statistics, in addition to advanced work in specialized areas. A graduate-level Program in Survey Methodology was established in 2001, seeking to train future survey methodologists in communication studies, economics, educations, political science, psychology, sociology, and statistics. The program offers a certificate, master of science, and doctorate degrees through the University of Michigan.

Jody Smarr

See also Consumer Sentiment Index; Joint Program in Survey Methods (JPSM)

Further Readings

ICPSR Summer Program in Quantitative Methods of Social Research: http://www.icpsr.umich.edu/training/summer
Institute for Social Research: http://www.isr.umich.edu
Summer Institute in Survey Research Techniques: http://www.isr.umich.edu/src/si

Institutional Review Board (IRB)

Institutional review boards (IRBs) are committees charged with the review and monitoring of research (including surveys) involving human participants. The basic principles of human research protection used today in the United States were outlined in the *Nuremberg Code* and were developed in response to the Nazi atrocities. Voluntary informed consent to research participation is at the core of that code. In response to research participant abuse in the first half of the 20th century, IRBs were mandated in the United States by the Research Act of 1974. In 1978, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research submitted the *Belmont Report*, which recommended the basic ethical principles underlying the acceptable conduct of research involving human participants. Those principles are (a) respect for persons (autonomy), (b) beneficence, and (c) justice. These basic principles continue to form the foundation for ethical conduct of research involving human participants.

Since 1974, all research funded by the Department of Health and Human Services has been required to undergo review to ensure ethical treatment of research participants. Most organizations receiving federal funding have assurance agreements that extend this ethical review to all research conducted by the institution. Although not yet enacted, in recent years, several federal legislators have expressed interest in expanding these protections to all public and private research—irrespective of funding source.

The constitution and utilization of IRBs were included in the regulations that were codified in Title 45 (Public Welfare—Department of Health and Human Services) *Code of Federal Regulations* (C.F.R.) Part 46 (Protection of Human Subjects). The regulations require that IRBs have a minimum of five members of varying experience and expertise as well as broad demographic representation. At least one member must have scientific expertise and at least one member must be from a nonscientific area. One member of the IRB must be from outside the institution. IRBs are also allowed to invite individuals with special expertise or knowledge to provide consultation and information on individual protocols, where needed.

Review of research activities typically involves the submission of a detailed overview of the research protocol utilizing a submission form. The form asks specific questions regarding how participants will be recruited, what they will be asked to do, details of the research design, and how the data will be transmitted, stored, and disseminated. The IRB reviews this information and performs a risk–benefit analysis to assure that any risks to the participants are offset by benefits to the participants or society. The review includes the following seven components: (1) identification of risks associated with the research participation; (2) identification of probable benefits of the research; (3) to the extent possible, assurance that risks are minimized; (4) determination of whether risks are proportionate to benefits; (5) assurance that participants are given accurate and complete information about the potential risks and potential benefits; (6) determination of the adequacy of privacy and confidentiality protections; and (7) determination of the best review intervals and any necessary monitoring of data collection.

The IRB has the authority to approve a protocol, to disapprove a protocol, or to ask for revisions or modifications of a protocol before approving the protocol. The IRB's decision and any required modifications are made in writing to the investigator. If a project is disapproved, the investigator is notified of the rationale and is provided with an opportunity to respond to the IRB regarding the disapproval.

Many protocols require review by the fully convened IRB; however, most minimal risk research receives an expedited review (conducted by the IRB chair or his or her designee) if it falls into one of seven categories defined by the regulations. Other minimal risk research may be reviewed initially and then granted an exemption from review if it meets specific criteria defined by the regulations and the local IRB. The determination of whether or not a research project meets the definition for exempt or expedited review is discretionary on the part of the IRB. The application of the various types of review is determined by the local IRB. At most academic institutions, for example, all research is reviewed by the IRB to assure that basic ethical standards are met. Minimal risk surveys are typically reviewed under an exempt or expedited category.

The IRB typically requires documentation of informed consent (e.g., a form signed by the participant that outlines the project, the risks, and benefits). However, the IRB may also approve a waiver of documentation of consent. This often is done for survey projects where obtaining a signature would not be possible or feasible or necessary in light of the minimal risks involved. For example, in a random-digit dialing telephone interview, consent is obtained (typically in oral mode), but no signature is required.

The IRB must review projects at least annually (for those lasting more than 1 year) but may require a shorter interval for projects that are more than minimal risk. The IRB is also authorized to have one or more members observe the recruiting, consent, and research process or may enlist a third party to observe to ensure that the process meets the desired ethical standards and desired levels of risk and confidential treatment of data. A protocol approved by an IRB may still be disapproved by the institution for some other reason. However, if a protocol is disapproved by the IRB, it may not be approved by the institution.

Mary E. Losch

See also Ethical Principles; Informed Consent; Protection of Human Subjects; Voluntary Participation

Further Readings

Citro, C., Ilgen, D., & Marrett, C. (Eds.). (2003). *Protecting participants and facilitating social and behavioral sciences research*. Washington, DC: National Academies Press. Retrieved May 2003 from http://www.nap.edu/books/0309088526/html

Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart A), 101–124.

Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart B), 201–211.

Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart C), 301–306.

Federal Policy on the Protection of Human Subjects, 45 C.F.R. 46 (Subpart D), 401–409.

U.S. Department of Health and Human Services. (1993). *Institutional review board guidebook* (chap. 3). Retrieved March 29, 2008, from http://www.hhs.gov/ohrp/irb/irb_chapter3.htm

Interaction Effect

An interaction effect is the simultaneous effect of two or more independent variables on at least one dependent variable in which their joint effect is significantly greater (or significantly less) than the sum of the parts. The presence of interaction effects in any kind of survey research is important because it tells researchers how two or more independent variables work together to impact the dependent variable. Including an interaction term effect in an analytic model provides the researcher with a better representation and understanding of the relationship between the dependent and independent variables. Further, it helps explain more of the variability in the dependent variable. An omitted interaction effect from a model where a nonnegligible interaction does in fact exist may result in a misrepresentation of the relationship between the independents and dependent variables. It could also lead to a bias in estimating model parameters.

As a goal of research, examination of the interaction between independent variables contributes substantially to the generalization of the results. Often a second independent variable is included in a research design, not because an interaction is expected, but because the absence of interaction provides an empirical basis for generalizing the effect of one independent variable to all levels of the second independent variable.

A two-way interaction represents a simultaneous effect of two independent variables on the dependent variable. It signifies the change in the effect of one of the two independent variables on the dependent variable across all the levels of the second independent variable. Higher-order interactions represent a simultaneous effect of more than two independent variables on the dependent variable. Interaction effects may occur between two or more categorical independent variables as in factorial analysis of variance designs. It may also occur between two or more continuous independent variables or between a combination of continuous and categorical independent variables as in multiple regression analysis. To illustrate these effects, the following sections start with the interpretation of an interaction effect between two categorical independent variables as in ANOVA, followed by the interpretation of an interaction effect between one categorical independent variable and one continuous independent variable, and finally the interpretation of an interaction effect between two continuous independent variables. Interpretation of three or more higher-order interaction terms effects follow the same logic of interpreting the two-way interaction effect.

Interaction Between Two Categorical Independent Variables

Consider a survey research study investigating the effectiveness of incentives and postal mailers on response rates in a mail survey. Incentive amount is the first categorical independent variable (A) with three groups; A_1 is a control group who receives no incentive, A_2 is a \$1 incentive group, and A_3 is a \$5 incentive group. The second categorical independent variable (B) is type of mailer with B_1 for First-Class Mail and B_2 for Federal Express. Response rates to the mail survey in percentages (Y) are the dependent variable for the study. In this typical (3×2) ANOVA there is (a) a possible main effect for incentive amount on response rates, (b) mailer type as a possible main effect on response rates, plus (c) a possible interaction effect between incentive amount and mailer type on response rates. A significant interaction effect suggests that the differences in the effects of incentive amount on response rates depend on mailer type (and vice versa). That is, in this example, the average differences in incentive amount effect on response rates are different in magnitude and possibly in direction for First Class versus what they are for Federal Express. Conversely, one can say that the average difference (in magnitude and direction) for response rates between a First-Class envelope and a Federal Express envelope depends on the incentive amount a household receives. An insignificant (negligible) interaction effect between incentive amount and mailer type on response rates suggests that the differences in response rates across incentive amounts are essentially the same for First Class and Federal Express. One can also interpret an insignificant interaction effect between mailer type and incentive amounts by recognizing that the difference between First Class and Federal Express is basically the same (in magnitude and direction) across the three incentive amounts.

Graphing interaction effects often simplifies their interpretation. It also helps analysts identify the two types, ordinal and disordinal interaction effects (as shown in Figures 1 and 2).

Ordinal Interaction

Ordinal interaction occurs when the magnitude of the differences between the levels of one independent variable on the dependent variable varies across the levels of the other independent variable, yet the direction (order) of these differences stays the same. In the response rate example, a case of an ordinal interaction might occur when, for example, the difference in response rates between First Class and Federal Express varies for each incentive amount, yet Federal

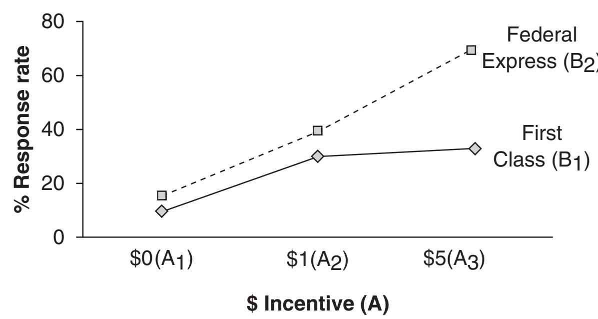

Figure 1 Ordinal interaction

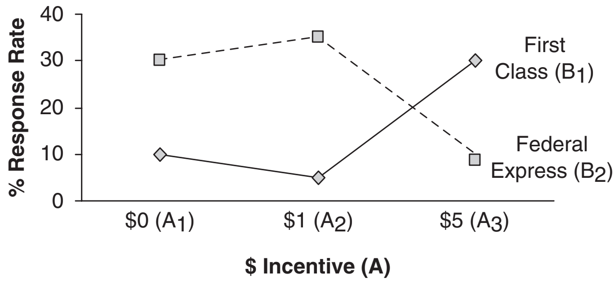

Figure 2 Disordinal interaction

Express always leads to a higher response rate than First Class, across all incentive amounts that are tested in the experimental design. Figure 2 illustrates the constant (consistent) order in the difference between Federal Express and First Class, where Federal Express (B_2) always has a higher response rate than First Class (B_1) for every incentive amount tested.

Disordinal Interaction

A disordinal interaction (as shown in Figure 2) occurs when the difference between the levels of one independent variable on the dependent variable varies in magnitude and direction across the levels of the other independent variable. An example of a disordinal interaction would occur if Federal Express leads to higher response rate when \$0 and \$1 are sent as incentives, but First Class leads to higher response rates when \$5 is sent. This type of inconsistent interaction is depicted in Figure 2.

No Interaction

No interaction effect exists when the differences in response rates between types of mailers are of the same magnitude and same order across all three incentive amounts. This case results in having parallel lines for First Class and Federal Express, as depicted in Figure 3.

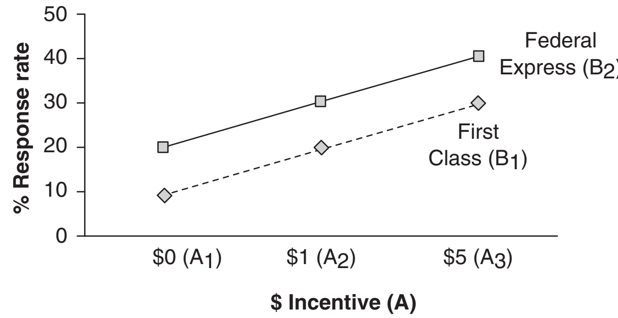

Figure 3 No interaction

Interaction Between One Categorical Variable and One Continuous Variable

Studying the interaction effect of continuous and categorical independent variables on a dependent variable is usually done through a regression analysis. A categorical independent variable usually is created using one of several coding methods when it is used in a regression analysis. One commonly used coding method is a *dummy coding*, where $C - 1$ dummy variables (taking on the values of 0 or 1) are created to represent the C categories of the variable with one reference category (e.g., a categorical variable such as race might be coded so that reference category is White and three dummy variables are created: Asian, Black, and Other race). The estimated regression coefficient for a dummy variable represents the average difference in the dependent variable between the category of the dummy variable and the reference category. The coefficient associated with the continuous independent variable represents the effect of the independent variable on the dependent variable.

Consider a study surveying people's attitudes toward government war policy and how it might be influenced by their political party affiliation (Republican versus Democrat) and the time they spend watching the Fox News TV channel. Party affiliation constitutes the categorical independent variable in this example. Time (T) in minutes spent watching the Fox News TV channel is the continuous independent variable.

To model the effect of party affiliation on attitude through a regression analysis we create a dummy variable (D) with Democrats coded "1" and Republicans coded "0."

Equation 1 provides the coefficients (b's) for a full estimated regression model that predicts people's attitude (Y) from their party affiliation (D), the time they spend watching Fox News (T), and a product term between party affiliation and time ($T \times D$). These coefficients represent the main effect of party type, main effect of time spent watching the Fox News TV channel, and an effect for the interaction on people's attitudes toward government war policies.

$$\hat{Y} = b_o + b_1\text{D} + b_2\text{T} + b_3(\text{D x T}) \quad (1)$$

Substituting the values of the variable D in Equation 1 generates two regression equations with different watching time effect on people's attitudes on war policy, one for the Republicans and the other for the Democrats. The two main effects are reflected in the coefficients b_2 and b_3 representing the interaction effect between time and party affiliation on people's attitude.

$$\hat{Y} = b_o + b_2\text{T} \quad (2)$$

$$\hat{Y} = b_o + b_1 + (b_2 + b_3)\ \text{T} \quad (3)$$

The value of b_2 in Equation 2 reflects the effect of time Republicans spend watching Fox News TV channel on their attitude. The sum of b_2 and b_3 in Equation 3 is the effect of time Democrats spend watching Fox News TV channel on their attitude. A significant test of b_3 would signify a nonnegligible interaction effect of party affiliation with the time spent watching the Fox News TV channel on people's attitude toward war policy.

Interaction Between Two Continuous Independent Variables

An interaction effect between two continuous independent variables on a dependent variable is expressed as a multiplicative term in a multiple regression analysis. A full regression model that predicts, say, parents' attitude toward schools (Y) from schools' involvement in the community (SIC) and parents' social economic status (SES) as two continuous independent variables and their interaction (SIC × SES) is presented in Equation 4.

$$\hat{Y} = b_o + b_1\text{SIC} + b_2\text{SES} + b_3(\text{SIC x SES}) \quad (4)$$

The value of b_3 represents the interaction effect between SIC and SES on parents' attitude, Y. It reflects SIC effect on parents' attitude, Y, conditioned on the levels of SES. To understand the interpretation of the interaction effect between SIC and SES, it is better to reorganize Equation 4 into

$$\hat{Y} = b_o + b_2\text{SES} + (b_1 + b_3\ \text{SES})\text{SIC}. \quad (5)$$

Equation 5 reveals that the effect of SIC, $(b_1 + b_3\text{SES})$, on parents' attitude (Y) depends on the levels of SES. One must also realize that Equation 4 can also be rearranged to express the same interaction between SES and SIC as the conditional effect of the SES variable on the dependent variable Y, conditioned on the values of the SIC variable. A close look at Equation 5 reveals its similarity to Equation 3. In fact, if SES is categorized into two categories such as high SES and low SES, Equation 5 becomes equivalent to Equation 3.

A nonsignificant value of b_3 implies a negligible or no interaction effect between the two variables SES and SIC. For a nonsignificant interaction, the regression model simplifies to

$$\hat{Y} = b_o + b_1\text{SIC} + b_2\text{SES}. \quad (6)$$

The coefficient b_1 represents the same effect of school involvement in the community on parents' attitude across all levels of parents' social economic status. Similarly b_2 represents the same effect of parents' social economic status on their attitude across all the levels of school involvement in the community.

Rafa M. Kasim

See also Analysis of Variance (ANOVA); Dependent Variable; Experimental Design; Factorial Design; Independent Variable; Regression Analysis; Research Design; Research Question

Further Readings

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Kutner, M., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models*. Boston: McGraw-Hill Irwin.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. Belmont, CA: Wadsworth.

Interactive Voice Response (IVR)

Interactive voice response (IVR) is a data collection technology that uses a recorded voice to ask survey questions by telephone, in place of live interviewers. Respondents enter their answers by pressing the buttons on the keypad of their touchtone telephone. An IVR system controls the presentation of the survey questions, captures the responses entered via touchtone, prompts respondents to answer questions, and offers automated help to respondents. IVR is also known as telephone audio computer-assisted self-interviewing (T-ACASI) and touchtone data entry. These terms all refer to computerized telephone data collection systems where respondents answer survey items via automated self-administered procedures, as opposed to giving verbal answers to live interviewers.

IVR has two primary uses in survey research. First, IVR can be used to replace interviewer-administered telephone data collection. Potential respondents may be contacted first by telephone interviewers and then switched to IVR, or they can be contacted by another mode (e.g., mail or Internet) and provided a call-in number to use for completing an IVR interview. Another use of IVR is to provide telephone survey respondents greater privacy in responding to questions of a potentially sensitive nature. When used for this second purpose, IVR interviews are typically conducted through initial contact from telephone interviewers who then switch respondents to IVR for the sensitive items. The respondent is then switched back to the interviewer after answering the sensitive questions. Regardless of the specific purpose, IVR data collection typically involves a relatively short, simple interview or a brief module that is part of a longer interview.

IVR offers a number of potential advantages over interviewer-administered modes of telephone data collection. First, because a pre-recorded voice is employed to administer all survey items, IVR respondents are all read questions, response options, and instructions in the exact same way. This provides a higher degree of interview standardization than interviewer-administered telephone data collection, where interviewers' vocal qualities, reading skills, and presentation skills vary.

A second advantage of IVR is providing greater privacy for answering questions, especially survey items of a potentially sensitive nature that could be affected by socially desirable reporting. Because IVR respondents enter their answers by pressing touchtone buttons, they do not have to be concerned about giving their responses to a live interviewer or about others (e.g., family members) hearing their responses. Research indicates respondents in IVR mode give more nonnormative answers to questions of a sensitive nature compared to respondents in computer-assisted telephone interviewing (CATI) mode. Examples include greater reports of illicit substance use, certain sexual behaviors, and negative satisfaction ratings.

Third, when IVR surveys are conducted with no interviewer involvement, this mode can be a much more cost-effective method than CATI. This version of IVR data collection saves on both the costs of recruiting, training, and supervising interviewers as well as telecommunication costs. Because telephone charges are incurred only when potential respondents call in to complete IVR interviews, these costs are lower than costs in the interviewer-administered mode where multiple outbound calls are typically made to contact each sampled unit.

IVR does have some potential limitations compared to interviewer-administered modes of telephone data collection. First, IVR mode provides unique opportunities for unit nonresponse or incomplete interviews. When respondents are left to complete an IVR interview on their own time, they are less likely to participate without the motivation provided by interviewers. Similarly, in surveys where interviewers first recruit respondents and then switch them to IVR mode, respondents have an opportunity to terminate the interview at that point, without being exposed to the persuasive efforts of an interviewer.

Second, concerns about respondent patience and terminations limit the length and complexity of IVR surveys. More lengthy IVR surveys introduce considerable risk that respondents will terminate interviews prior to completion, because no interviewer is present to encourage continued participation. Complicated surveys may also increase the risk of respondents either terminating the interview or providing inaccurate responses or no responses to some items to reduce the survey burden.

Third, the lack of an interviewer creates a potential risk of measurement error in IVR interviews, as in other self-administered modes of data collection. Without an interviewer to clarify response tasks, probe inadequate responses, and record answers accurately, IVR data can introduce other sources of respondent-related measurement error that are not as likely to occur in interviewer-administered surveys. Although IVR systems can provide a help button for respondents to use when they are having difficulty completing a question, respondents may be hesitant to use this option when they need it or they may not receive sufficient help to accurately answer an item.

Finally, when IVR data collection is used as part of an interviewer-administered survey where interviewers recruit and screen eligible subjects, cost savings compared to CATI data collection mode may not be realized. Adding an IVR module to a CATI survey would actually add costs related to the additional programming and management required for using these two modes. Costs for recruiting, training, and supervising interviewers would be similar to those costs in CATI-only mode.

Douglas B. Currivan

See also Audio Computer-Assisted Self-Interviewing (ACASI); Computer-Assisted Telephone Interviewing (CATI); Privacy; Respondent-Related Error; Sensitive Topics; Social Desirability; Touchtone Data Entry

Further Readings

Cooley, P. C., Miller, H. G., Gribble, J. N., & Turner, C. F. (2000). Automating telephone surveys: Using T-ACASI to obtain data on sensitive topics. *Computers in Human Behavior*, *16*, 1–11.

Tourangeau, R., Miller, D., & Wilson, D. (2002). Self-administered questions by telephone: Evaluating interactive voice response. *Public Opinion Quarterly*, *66*, 265–278.

Villarroel, M. A., Turner, C. F., Eggleston, E., Al-Tayyib, A., Rogers, S. M., Roman, A. M., et al. (2006). Same-gender sex in the United States: Impact of T-ACASI on prevalence estimates. *Public Opinion Quarterly*, *70*, 166–196.

Intercoder Reliability

Intercoder reliability refers to the extent to which two or more independent coders agree on the coding of the content of interest with an application of the same coding scheme. In surveys, such coding is most often applied to respondents' answers to open-ended questions, but in other types of research, coding can also be used to analyze other types of written or visual content (e.g., newspaper stories, people's facial expressions, or television commercials). Intercoder reliability is often referred to as *interrater* or *interjudge reliability*. Intercoder reliability is a critical component in the content analysis of open-ended survey responses, without which the interpretation of the content cannot be considered objective and valid, although high intercoder reliability is not the only criteria necessary to argue that coding is valid. Intercoder reliability is a standard measure of research quality, and a low level of intercoder reliability may suggest weakness in coding methods, including poor operational definitions with unclear coding categories and poor coder training.

Although there are more than 30 different statistical measures or indices of intercoder reliability, only a handful of measures are widely used and there is no consensus on the single best measure. Among all, for its simplicity and ease of use, *percent agreement* is the single most widely used index. It is measured by the proportion of coding decisions that reached agreement out of all coding decisions made by a pair of coders. However, it does not account for agreement that could occur by chance. The proportion of agreement by chance alone is higher when fewer coding categories are given and lower when more coding categories are given. As the number of coding categories increase, however, high percent agreement becomes more difficult to achieve. Percent agreement is also limited to nominal coding with only two coders with the same number of coded units. Holsti's *CR*, a variation of percent agreement index, accounts for different numbers of coded units each coder produces. Scott's pi (π), on the other hand, takes into account the agreement that can occur by chance. It also accounts for the number of coding categories and distribution of coded categories. Scott's pi (π), however, is appropriate only for nominal level coding. Cohen's kappa (κ) is another widely used index that also accounts for chance agreement. Kappa is commonly used for the nominal level coding of behaviors. Although this index has been adapted to situations where multiple coders evaluate a different set of units, it is known to be most appropriate when one pair of coders judges the same set of units without violating the assumptions of independent coders and random coder errors. A more flexible measure, although more demanding mathematically, is Krippendorff's alpha (α). This index can account for chance agreement, different levels of coding (i.e., nominal, ordinal, interval, and ratio), multiple coding categories, different sample sizes, and missing data. Generally, the selection of proper index will depend on the levels of coding, number of coded categories if coded nominal, number of coders, and number of coded units. Indices that measure association (i.e., chi-square), internal consistency (i.e., Cronbach's alpha), or correlation (i.e., Pearson's r) should not be used for the purpose of measuring intercoder agreement simply because they do not measure the level of agreement between coders. For example, if one coder consistently codes 1 point higher on a 5-point scale than the other coder, they are completely correlated (Pearson's $r = 1$) or consistent (Cronbach's alpha $= 1$), but the agreement (Krippendorff's alpha) is about half of that when treated as interval level coding with a pair of 10 coded units.

Intercoder reliability coefficients range from 0 (complete disagreement) to 1 (complete agreement), with the exception of Cohen's kappa, which does not reach unity even when there is a complete agreement. In general, coefficients .90 or greater are considered highly reliable, and .80 or greater may be acceptable in most studies. When coder agreement is poor, it can be improved by training coders with coding schemes that define coding categories as clearly as possible before the actually coding and with a couple of practice runs with small subsamples of cases to refine the coding scheme.

Young Ik Cho

See also Coding; Content Analysis; Cronbach's Alpha; Open-Ended Question

Further Readings

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

Neuendorf, K. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

INTERNAL VALIDITY

As explained in the 1960s by Donald Campbell and Julian Stanley in their seminal book *Experimental and Quasi-Experimental Designs for Research*, internal validity refers to the extent to which the methodological research design used by a researcher can provide empirical evidence to test the possible cause-and-effect relationship between an independent variable (the antecedent), X, and a dependent variable (the consequence), Y. Without adequate internal validity, researchers may offer logical, reasoned arguments to speculate about the possible causal nature of any correlational relationship they observe between X and Y in their data, but they cannot use the internal strength of their research design to bolster such reasoning. Thus, although researchers can (and routinely do) draw casual inferences about "X causing Y" based on speculation from the results of many types of nonexperimental research designs, it is only with carefully planned experiments and quasi-experiments that researchers can draw internally valid conclusions about cause-and-effect relationships with a high degree of confidence.

Unfortunately, many survey researchers have not received training in experimental design and do not appear to appreciate the importance of planning research that can yield internally valid findings. As such, they often fail to take full advantage of deploying controlled experimentation with random assignment in their surveys. The following example, which occurred in 2000, helps illustrate this point.

Example of a Survey Lacking Internal Validity

The results of a nonexperimental pre-election study on the effects of political advertising were posted on the American Association Public Opinion Research listserve (AAPORnet) in 2000. This survey-based research was conducted via the Internet and found that a certain type of advertising was more persuasive to potential voters than another type. By using the Internet as the data collection mode, this survey was able to display the ads—which were presented as digitized video segments—in real time to respondents/subjects as part of the data collection process and, thereby, simulate the televised messages to which voters routinely are exposed in an election campaign. Respondents were shown all of the ads and then asked to provide answers to various questions concerning their reactions to each type of ad and its influence on their voting intentions. This was done in the individual respondent's own home in a room where the respondent normally would be watching television. Here, the Internet was used very effectively to provide mundane realism to the research study by having survey respondents react to the ads in a context quite similar to one in which they would be exposed to real political ads while they were enjoying a typical evening at home viewing television. Unlike the majority of social science research studies that are conducted under conditions far removed from real life, this study went a long way toward eliminating the potential artificiality of the research environment as a serious threat to its overall validity.

Another laudable design feature of this study was that the Internet sample of respondents was chosen with a rigorous scientific sampling scheme so that it could reasonably be said to represent the population of potential American voters. The sample came from a large, randomly selected panel of households that had received Internet technology (*WebTV*) from the survey organization in their homes. Unlike most social science research studies that have studied the effects of political advertising by showing the ads in a research laboratory setting (e.g., a centralized research facility on a university campus), the overall validity of this study was not threatened by the typical convenience sample (e.g., undergraduates "volunteering" to earn course credit) that researchers often rely upon to gather data. Thus, the results of this Internet research were based on a probability sample of U.S. households and, thereby, could reasonably be generalized to the potential U.S. electorate.

As impressive as these features of this research design were, the design had a serious, yet unnecessary, methodological flaw—one that led the researchers to miss a golden opportunity to add considerably to the overall validity of the conclusions that could have been drawn. The research design that was used displayed all the political ads to each respondent, one ad at a time. There were no features built into the design that controlled either for the possible effects of the order in which the respondent saw the ads or for having each respondent react to more than one ad within the same data collection session. As such, the cause-and-effect conclusions that could be drawn from this nonexperimental study design about which ads "caused" stronger respondent reactions rested on very weak methodological footing. Since no design feature was used to control for the fact that respondents viewed multiple ads within the same data collection session, the conclusions drawn about the causality underlying the results remained little more than speculations on the part of the researchers, because such factors as the order of the ads and the number of ads were not varied in a controlled manner by the researchers. Unfortunately this missed opportunity to use an experimental design is all too common in many survey-based research studies in the social sciences.

This study perfectly lent itself to the use of various experimental designs whereby a different political ad (i.e., the experimental stimuli), or different subsets of ads, could have been randomly assigned to different subsamples of respondents. Or, the order of the presentation of the entire set of political ads could have been randomly assigned across respondents. In either case, an experimental design with random assignment would have provided the researchers a far stronger basis (one with greater internal validity) from which to draw their causal inferences. Furthermore, such an experimental approach would have had little or no cost implications on the research budget, and a design where one and only one ad was shown to any one respondent would likely have saved data collection costs and yet would have been more powerful in supporting causal interpretations.

Conducting Internally Valid Survey Research

Cook and Campbell define internal validity as the approximate validity with which one infers that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause. There are three conditions for establishing that a relationship between two variables (X and Y) is a causal one, as in "X causes Y." The researcher must demonstrate

1. Covariation, that there is a reliable statistical relationship between X and Y.
2. Temporal order, that X occurs before Y occurs.
3. An attempt to eliminate other plausible explanations than changes in X for any observed changes in the dependent variable (Y). The use of a true experimental design with random assignment of respondents to different levels of X is of special value for this last condition to be met.

The simple formula, $Y = f(X)$, often is used to depict covariation between two variables, X and Y, and is read, "Y is a function of X." The concept of internal validity essentially addresses the nature of the equal sign (=) in the equation; that is, is the relationship between X and Y a causal one? For internal validity to exist in a research study there must be covariation demonstrated between X and Y, and, therefore, X must predict Y to some statistically reliable extent. But the equal sign (and the covariation it implies) in itself does not provide internally valid evidence that a cause-and-effect relationship between X and Y has been demonstrated.

If the relationship in the formula, $Y = f(X)$, is causal, then it presupposes that X precedes Y in a temporal sense. This is what distinguishes this specification from one that says Y is the cause of X, or that each causes the other, or that each is caused by some other unspecified variable (Z)—any of which could be the interpretation of the observed correlation between two variables. Only by the use of a controlled experiment or series of experiments can the nature and direction of these interrelationships be parceled out through the implementation of the independent variable (under the control of the researcher) followed by the measurement of the dependent variable. Only in this way can evidence be gathered to demonstrate with confidence that the relationship is a causal one and not merely one of noncausal covariation.

The essential design feature of an experiment is the use of random assignment of respondents to different experimental conditions. The logic here is that with random assignment of different respondents to different conditions (i.e., the different levels of the independent variable, X), all other factors will be equivalent except for the differences that the researcher directly controls in implementing the independent variable of interest. If, for example, statistically significant differences in the mean of the dependent variable (Y) then are found between the randomly assigned groups, these differences then can be attributed to the levels of the independent variable that the researcher has controlled. Then the researcher usually will have a solid basis to conclude that it was the controlled differences in the independent variable (X) that *caused* the observed differences in Y between the groups.

A simple survey-based example of this occurs when a group of respondents are randomly assigned to two conditions: often called a *control* condition and a *treatment* condition. For this example, assume a questionnaire employs a so-called split-half experiment, whereby one random half of respondents are exposed to the standard wording of a question (e.g., the wording used in the 2000 Census to measure whether someone is Hispanic: "Are you Spanish, Hispanic, or Latino?"). The group receiving this standard wording is the control group. The other random half of respondents would be asked the question with some altered version of the wording (e.g., "Are you of Spanish, Hispanic, or Latino origin?"). In this example, the group of respondents seeing or hearing the word *origin* in their questionnaire is the treatment group (sometimes called the experimental group), as they are receiving a different wording of the question to investigate what effect adding the word *origin* will have on the proportion in the treatment group that answers "Yes." The researcher has controlled the administration of the question wording or X, the independent variable, in order to learn whether the change in wording causes a change in Y, the dependent variable, that is, the proportion of people who say "Yes" (i.e., they are Hispanic). This control over administration of the independent variable (X) is exercised via random assignment so that, in theory, nothing else is dissimilar between the two groups except for the slight change of wording between the control question and the treatment question. Thus, random assignment is the equivalent to holding "all other things equal." Because of the strong internal validity of the experimental design in this example, the researcher can conclude with great confidence that any statistically significant difference between the two groups in the proportion that answered, "Yes" to being Hispanic is associated with (i.e., caused by) the presence or absence of the word *origin* (the independent variable).

Threats to Internal Validity

To better appreciate the power of an experimental design and random assignment, it is worth a brief review of some of the major reasons that cause-and-effect inferences drawn from *nonexperimental* research lack internal validity and therefore are subject to many threats to their overall validity.

Selection

Too often the selection of the respondents that constitute different comparison groups turns out to be the main threat to a study's internal validity. For example, if a survey researcher sampled two different municipalities and measured the health of residents in each community, the researcher would have no statistical or methodological grounds on which to base any attributions about whether living in one community or the other *caused* any observed differences between the average health within the respective communities. In this example, no controlled effort was built into the study to make the two groups equivalent through random assignment, and of course, in this example that would not be possible. As such, any observed differences between the health in one community versus another could be due to countless other reasons than place of residence, including a host of demographic and behavioral differences between the residential populations of each community.

Thus, any time two (or more) groups have been selected for comparison via a process other than random assignment, the researchers most often will have no solid grounds on which to draw valid inferences about what may have caused any observed difference between the two groups. (An exception is the possibility that a researcher deployed a quasi-experimental design, one without true random assignment, but one that may have had other rigorous design features that avoided some of the potential threats to internal validity).

Unfortunately, not using an experimental design does not stop many researchers from making unfounded causal attributions. This is especially the case in the field of medical research when the health of large samples of volunteer patients is tracked over time. Such a panel survey may find many significant correlations between behavior and health (e.g., eating a lot of carrots is associated with better eyesight), but this is mere covariation, and the study design, with its lack of random assignment to comparison groups, provides no internally valid basis to support causal inferences between the measured behaviors and health.

Furthermore, any study that allows respondents to self-select themselves into different comparison groups will suffer from selection as a threat to its internal validity. However, this point is different from the common way that self-selection into a sample is thought of. So-called convenience samples suffer from a self-selection sampling bias, as the researcher has no means of knowing whether a larger population is represented by a self-selected sample (this is an issue that affects external validity, i.e., its generalizability). However, the researcher could legitimately build a valid experiment into a survey that uses a convenience sample, simply by randomly assigning the self-selected respondents to different comparison groups. Thus, as long as the respondents do not self-select themselves into the treatment and control groups, the internal validity of the study is not threatened, even if they have self-selected themselves into the larger sample.

History

This potential threat to internal validity refers to the possibility that something other than the independent variable may have taken place between the time respondents were first exposed to the independent variable and time of the measurement of the dependent variable. If so, then a differential history effect may have caused any observed differences among respondent groups in the dependent variable.

To illustrate this, consider a survey of attitudes toward local police being administered in two different, yet socioeconomically similar, communities to establish a baseline. Imagine that the survey found that these two communities held essentially similar pretest attitudes. Then imagine that in one of the two communities, local police implemented a foot patrol program putting many more police officers on neighborhood streets. After this program has been implemented for several months, both communities are then resurveyed and the community with the new foot patrol program is now found to hold significantly more positive attitudes than the other community.

Could the researchers conclude with confidence (i.e., with strong internal validity) that the foot patrol program caused the improvement in attitudes? The answer is “No” for many reasons, including that there was no way for the researcher to control for whatever else may have occurred locally between the time that the two pretest surveys were conducted that may have led to the attitudes in one community to change compared to the other. For example, a major crime may have been solved in one community in the intervening period. Was this the cause of more positive attitudes toward the police, or was it the foot patrols? This is how the differential history of two groups can confound any interpretation of cause when a true experiment is not used.

Furthermore, even if a research study starts out as a true experiment, subsequent uncontrolled history between randomly assigned groups can undermine the experiment and, thereby, undermine its internal validity. For example, imagine a study in which the interviewers at a survey organization were randomly assigned into two groups to be trained, separately, to administer one of two different introductory spiels to randomly selected households in order to determine the differential effects on response rates of the two introductions. If something eventful happened at one of the training sessions other than the difference in the content related to the respective introductory spiels—for example, an interviewer and the trainer got into a heated argument about the wording of the introductory spiel, thereby lowering the confidence of the rest of the interviewers in that group regarding the effectiveness of that introductory spiel—then this differential history could pose a serious threat to the internal validity of this research study, despite it being originally designed as a true experiment. If this were to happen, then the researchers would have a weakened basis on which to conclude that it was the content of the different introductions *and only that content* that caused any observable differences in response rates between the two groups of interviewer and their respondents.

All this notwithstanding, in many survey-based experiments, history is not a likely threat to internal validity because the dependent variable often is gathered immediately after the administration of the independent variable (e.g., most wording experiments built into a questionnaire require that the respondent answer the question immediately after being exposed to the wording), but in other instances the researcher must be very conscious of the possibility that history may have undermined the integrity of the experimental design.

Instrumentation

Anytime a measurement instrument, for example, a survey question, is changed between a pre- and post-period, any observed changes in the dependent variable of interest may be due solely to the change in instrumentation as opposed to real changes between the two groups due to a treatment or stimulus. For example, take a panel survey with two waves of data collection in which all respondents were asked, *Do you support or oppose the President's new plan to reduce taxes?* in Wave 1 data collection. Suppose after Wave 1, a random half of the respondents were exposed to a direct mail campaign touting the popularity of the new tax-relief plan. Suppose also that after Wave 1, the President began actively campaigning on behalf of the new tax-relief plan and received consistently positive press coverage. After some passage of months, another wave of data is gathered from the same respondents, but using the following question: *Do you support or oppose the President's popular plan to reduce taxes?* Imagine that at Wave 2, a sizably larger proportion of respondents who were exposed to the direct mail campaign said they supported the plan than had supported it at Wave 1 and that this increase was larger than the increase in support among the nontreatment group. Would this mean that the direct mail campaign exposure caused the apparent growth within that portion of the sample exposed to it?

The answer is "No, not necessarily," because although the small change in the wording of the measure at Wave 2 may appear innocuous—and, given the positive press coverage, might appear to be an appropriate wording change—the use of the word *popular* in the Wave 2 version of the questions could by itself have prompted (i.e., caused) more people to "conform" with majority public opinion and say "Yes" to the question than otherwise would have happened had the exact Wave 1 wording been used. This could especially be true for the respondents exposed to the direct mail campaign. In particular, the treatment (the direct mail campaign) may have interacted with the wording change in the post-test question to cause the disproportionate shift in expressed support of the new tax plan among the group exposed to the mail campaign. Thus, it is possible that the change in support among the treatment group would have been no different in size than the change among the group that did not receive the direct mail campaign had the question wording not been altered.

Mortality

Imagine that an experimental test of a new remedial science curriculum is implemented so that a large random sample of inner-city high school students is randomly assigned to a treatment group or a control group. The control group does not receive the remedial curriculum. The treatment group receives the remedial instruction during a special 30-minute class held only for them at the end of the regular school day. After 6 months of being exposed daily to the remedial curriculum, the treatment group actually scores lower in science knowledge than does the control group. Does this mean that the curriculum actually caused the treatment group to do more poorly on their science knowledge test?

Although that is possible, imagine instead that receiving the remedial education curriculum caused more students in the treatment group to remain in school after 6 months because they were receiving the special attention. However, in the control group, more students dropped out of school during the ensuing 6 months, with students having the lowest knowledge of science being the ones most likely to drop out. In this case, differential mortality (or differential attrition) would render the two groups no longer equivalent when the comparison was made between each group's average science knowledge score after 6 months. As such, researchers must guard against respondent/subject mortality threatening the internal validity of their experiments. And, even if the researchers cannot foresee or control against differential mortality, the possibility that this might occur must be measured and its possible effects taken into account before one can interpret experimental results with confidence. In particular, any survey-based experiment in which the experimental treatment causes differential response rates, but the dependent variable is something other than the response rate, is subject to the effects of differential mortality.

There are other threats to the internal validity that may undermine a research design's ability to support cause-and-effect reasoning. However, by using a true experiment with random assignment, the researcher is on much firmer ground in making valid causal attributions than without an experimental design.

Internal Validity/Random Assignment Versus External Validity/Random Sampling

Now that it has been explained that random assignment is the cornerstone of experimentation and the establishment of internal validity of a research design, it is worth observing that many survey researchers and students new to the field appear to confuse random assignment with random sampling, or at least seem not to appreciate the distinction. Random sampling is very much a cornerstone of external validity, especially when it is done within the context of a probability sampling design. The beauty and strength of high-quality survey research is that a researcher often can meld *both* random assignment and random sampling, thereby having strong internal validity *and* strong external validity.

Researchers who use the survey mode of data collection typically are much more familiar with the science of sampling than they are with the science of experimentation. Although many of them may not have prior familiarity with the term *external validity*, they are very familiar with the principles underlying the concerns of external validity: If one wants to represent some known target population of interest accurately, then one best utilize a sampling design that (a) well represents that population via a properly constructed sampling frame and (b) uses a random probability sampling scheme to select respondents from it, thereby allowing one to generalize research findings from the sample to the population with confidence and within a known degree of sampling error. Within a total survey error framework, the avoidance of coverage error and nonresponse error are each a necessary condition for achieving strong external validity, and together they comprise the sufficient condition. Thus, survey researchers need to use sampling frames that fully cover the target population they purport to represent and need to achieve an adequate response rate that avoids meaningful nonresponse bias.

The linkage between internal validity and external validity concerns whether any cause-and-effect relationship that has been observed in a survey research experiment can be generalized beyond the confines of the particular sample (subjects/respondents) on which the test was conducted. For example, the field of psychology has a long and honored history of using experimentation with strong internal validity. However, it also has the well-known (and not so honorable) history of questionable external validity for too often using unrepresentative convenience samples of college undergraduates.

Conclusion

One cannot understate the importance of the research "power" that is afforded by an experimental design in allowing a researcher to test the causal nature of the relationship between variables with confidence (i.e., with strong internal validity). As noted earlier, this often can be done at little or no additional cost in the data collection process, and sometimes it can even save costs as it may reduce the amount of data that must be gathered from any one respondent.

Paul J. Lavrakas

See also American Association for Public Opinion Research (AAPOR); Convenience Sampling; Coverage Error; Dependent Variable; Differential Attrition; Experimental Design; External Validity; Independent Variable; Interaction Effect; Noncausal Covariation; Nonresponse Error; Random Assignment; Self-Selected Sample; Split-Half; Total Survey Error (TSE); Validity

Further Readings

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally.

Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Lavrakas, P. J., Courser, M. W., & Diaz-Castillo, L. (2005, May). *What a difference a word can make: New research on the differences between Hispanic "origin" and Hispanic "identity" and their implications*. Paper presented at the 60th annual conference of the American Association for Public Opinion Research, Miami Beach, FL.

Visser, P., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in personality and social psychology* (pp. 233–252). Cambridge, UK: Cambridge University Press.

Winer, B. J., Brown, D. R., & Michaels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

International Field Directors and Technologies Conference (IFD&TC)

The International Field Directors and Technologies Conference (IFD&TC) is a voluntary organization of practitioners of survey data collection for nonprofit organizations. Its 1993 charter describes the IFD&TC as providing "an opportunity for survey research personnel who are involved in the direction of survey field operations, the technological aspects of survey data collection, and their management to meet and exchange ideas and information at an annual conference." This entry describes the current focus and somewhat unusual characteristics of the IFD&TC as a professional organization, its membership and governance, and its origins and development from predecessor organizations.

The Conference

The intent of the IFD&TC (according to its charter) "is to provide ***informal*** [emphasis added] interaction between field director, field technology, and survey management personnel of a type not usually available in professional conventions or through professional journals. The sessions [are] informal and focus on work in progress or recently completed, and on exchanges of information, practices, and opinions on relevant subjects of common interest. Finished papers ready for publication, public distribution, or production in official formal proceedings are not required [and not encouraged]." Extensive time is provided for discussion during sessions and for casual interchange during lengthy breaks between sessions. Because the attendees generally do not represent organizations in competition with each other, the discussions also are unusually frank, open, and mutually sharing. These characteristics have fostered strong personal loyalty by regular attendees and a welcoming attitude toward those new to the field.

Presentations frequently focus on the practical aspects of survey data collection by personal interviews, telephone interviews, mail questionnaires, email, the Web, and by fax. Topics often include the following: (a) the hiring, training, and supervision of interviewers; (b) methods of improving response rates; and (c) the technical issues of designing and managing computer-assisted personal, telephone, Web, and multi-mode surveys.

The annual conference is typically held in May, immediately following the annual conference of the American Association for Public Opinion Research (AAPOR). The IFD&TC is typically held in the same city (or a nearby city) as AAPOR, starting with an informal party on a Sunday evening and continuing through Wednesday morning. For details about the conference, see the IFD&TC Web site. To limit conference size and to maintain the tradition of openness, candor, sharing, and informality, attendance restrictions are part of the charter. Attendance is limited to "persons involved in survey research who have responsibility for field direction, technological aspects, or management of survey data collection, and who are associated with academic institutions, government agencies, or other nonprofit organizations."

Speakers from profit-making organizations have been invited to attend sessions on topics of common interest. No organization may send more than 16 attendees, and these must be divided between field director and field technology interest areas. All attendees are encouraged to participate in the program as presenters, panelists, facilitators, or chairs, and so forth.

The 2007 conference had approximately 250 attendees—a conference size the organizers and members have found appropriate for a good mix of topics and informal interaction. While the bulk of the attendees are from the United States, major contingents generally attend from organizations in Canada, Great Britain, and the Netherlands, with less frequent attendance from other countries.

Membership and Governance

There are no dues or membership fees beyond conference registration. The active members consist of persons in attendance at each annual conference. The IFD&TC mailing list consists of those who have attended any of the past 10 conferences. Each year nearly half of the participants are first-time attendees, often people who are relatively new to survey data collection. They regard the conference as a learning experience and an opportunity to establish a network with more experienced members of the field.

The conference includes an open meeting at which the officers of the organization are elected and the sites and dates of future meetings are chosen. Those in attendance at that conference are the voting membership. The only officers are the two program chairs, one each for field directors and field technologies for the following 2 years, plus a treasurer. The organization is incorporated in the Commonwealth of Virginia.

History

In 1968, John Scott of the University of Wisconsin wrote a memo "A Suggestion for a Conference on Field Problems in Survey Research," which became the basis for the first Field Directors Conference (FDC), the first predecessor of the current IFD&TC. The FDC was held annually through 1988, with attendance growing from the low 20s to the mid 60s. It is significant in the history of professional organizations that FDC had 26 successful annual meetings with no written charter. At that time, each participating organization was allowed only two attendees plus two more on the program. In 1988, computer-assisted telephone interviewing had become a more frequent topic, and the conference length was extended to include a focus for those interested in the technical side of survey data collection. William Nicholls of the U.S. Bureau of the Census convened a meeting with seven other FDC regulars, and they wrote a charter for a Field Technologies Conference.

From 1989 through 1993, the FDC and FTC had separate programs and registrations and met for one and a half days each, consecutively, in the same locale. In 1993 a joint meeting of the attendees adopted a common charter, based on the FTC charter, and the IFD&TC as it presently exists was formed.

Shapard Wolf and William L. Nicholls

See also American Association for Public Opinion Research (AAPOR); Field Director

Further Readings

IFD&TC: http://www.ifdtc.org

IFD&TC. (1993). *IFD&TC charter* (as amended 1999). Retrieved October 22, 2007, from http://www.ifdtc.org/charter.htm

IFD&TC. (2007). *Past conferences*. Retrieved October 22, 2007, from http://www.ifdtc.org/past_conferences.htm

Scott, J. (1968). *A suggestion for a conference on field problems in survey research*. Retrieved October 22, 2007, from http://www.ifdtc.org/scott_memo.htm

Wolf, S. (1998). *A brief history: International Field Directors and Technologies (1993–present), Field Technologies (1989–1992), and Field Directors (1951–1992)*. Retrieved October 22, 2007, from http://www.ifdtc.org/history.htm

International Journal of Public Opinion Research (IJPOR)

The *International Journal of Public Opinion Research* (IJPOR) is an academic quarterly founded and owned by the World Association of Public Opinion Research (WAPOR) and published by Oxford University Press. Its first issue appeared in 1989. Seymour Martin Lipset (United States), Elisabeth Noelle-Neumann (Germany), and Robert M. Worcester (United Kingdom)—all former presidents of WAPOR—were the founding editors. The journal continued an earlier WAPOR project, the *International Journal of Opinion and Attitude Research*, published between 1947 and 1951. The current editors of IJPOR are Wolfgang Donsbach (Germany) and Michael W. Traugott (United States). The editors are appointed by the WAPOR Council for a 3-year term, which can be extended. One of the editors acts as managing editor and runs the editorial office.

The geographical composition of the team of editors reflects the intention of the journal's owner to represent public opinion research in an international scope. The leadership of WAPOR started IJPOR as an international alternative to *Public Opinion Quarterly*, which at the time was essentially the only academic journal in the field. IJPOR operates in support of the basic purposes of WAPOR while reserving editorial independence. As outlined in the Constitution of WAPOR, the association is committed to "(a) promote in each country of the world the right to conduct and publish scientific research on what the people and its groups think and how this thinking is influenced by various factors, (b) promote the knowledge and application of scientific methods in this objective, (c) assist and promote the development and publication of public opinion research worldwide, (d) promote international cooperation and exchange among academic and commercial researchers, journalists and political actors, as well as between the representatives of the different scientific disciplines." IJPOR is charged to serve these aims through the publication of scientifically sound and practically relevant research in the field of public opinion research and by promoting research from a wide range of countries and research that is based on comparative studies.

IJPOR operates on a blind peer review system. The journal has an editorial board with some 30 scholars and practitioners from a variety of countries, but the number of referees—representing more than 30 different countries—goes much beyond that list. Due to an increasing number of submissions, the rejection rate has risen to 70%. Because of the more specialized nature of its content (compared to the more general social science journals), IJPOR is found only in the second half of the impact factor rankings of relevant social science journals. It has, however, improved its position in recent years. Circulation stands at about 2,000 subscribers, a number that includes consortia subscribers.

Each issue today contains five sections: regular research articles, shorter research notes, "World Opinion" (with topical comparative public opinion data), book and journal article reviews, and news on WAPOR. When IJPOR was founded, the WAPOR Council and the founding editors wanted to establish a truly international journal that would give a forum to social scientists and polling practitioners from around the world. Like its owner association, WAPOR, the journal was meant to bridges gaps between different professional fields (academia, business, administration), different disciplines (political science, sociology, communications, psychology, to name but the most important ones), and between theory building and empirical findings. Thus, the journal's content usually offers articles and research notes on five areas: (1) theories about the dynamics of public opinion, (2) methodological problems and developments, (3) the role of the news media in public communication, (4) public opinion research as a social and political problem, and (5) public opinion data on topical issues. Much of the research published in IJPOR is of comparative nature. IJPOR is a truly international journal. After conducting a quantitative content analysis of 43 leading journals in the wider field of communication studies that claim to be international, a scholar recently concluded that, based on content and authors, the IJPOR is the only journal that can be classified as international.

However, there is still room for improvement. A content analysis of the first 11 volumes (1989–1999) conducted by the editors revealed that about half the authors work in the United States, and that many other regions, particularly Africa, Latin America, and Eastern Europe, are underrepresented. IJPOR is a journal mainly for the intersubjective evidence produced by social scientists. This is reflected in the proportions between empirical tests of public opinion theories (e.g., the spiral of silence) and purely theoretical, nonempirical deliberations. While 17% of the IJPOR articles were dedicated to the former, only 3% dealt with theory only. This focus shows also in the total number of empirical studies on the pages of IJPOR: Overall, three in four articles presented some kind of empirical research.

Most of the studies using empirical methods are based on surveys. Among these studies is a high proportion of trend or panel surveys. In addition, and given the emphasis on the news media's role for public opinion, many IJPOR articles are based on quantitative content analyses or on a combination of both surveys and content analyses. About one in ten articles are purely theoretical, conceptual, or normative, that is, without any references to empirical results.

Wolfgang Donsbach

See also *Public Opinion Quarterly* (POQ); World Association for Public Opinion Research (WAPOR)

Further Readings

Lauf, E. (2005). National diversity of major international journals in the field of communication. *Journal of Communication*, *55*(1), 139–151.

World Association of Public Opinion Research. (2003, April 15). *Constitution*. Retrieved March 29, 2008, from http://www.unl.edu/WAPOR/constitution.html

International Social Survey Programme (ISSP)

The International Social Survey Programme (ISSP) is a continuing, annual program of cross-national collaboration. It brings together pre-existing social science projects and coordinates research goals, thereby adding a cross-national perspective to the individual, national studies.

ISSP evolved from a bilateral collaboration between the Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) of the Zentrum für Umfragen, Methoden, und Analysen (ZUMA) in West Germany and the General Social Survey (GSS) of the National Opinion Research Center (NORC), University of Chicago. Both the ALLBUS and the GSS are replicating, time-series studies. The ALLBUS has been conducted biennially since 1980 and the GSS annually since 1972. In 1982 and 1984 the ALLBUS and GSS devoted a segment to a common set of questions.

Meanwhile, in late 1983 the National Centre for Social Research (NCSR; then known as Social and Community Planning Research), London, which was starting a social indicators series called the British Social Attitudes Survey, secured funds to further international collaboration. Representatives from ZUMA, NORC, NCSR, and the Research School of Social Sciences, Australian National University, organized ISSP in 1984 and agreed to (a) jointly develop topical modules dealing with important areas of social science, (b) field the modules as a 15-minute supplement to the regular national surveys (or a special survey if necessary), (c) include an extensive common core of background variables, and (d) make the data available to the social-science community as soon as possible.

Each research organization funds all of its own costs. There are no central funds. The merging of the data into a cross-national data set is performed by the Zentralarchiv für Empirische Sozialforschung, University of Cologne, in collaboration with the Análisis Sociológicos, Económicos y Polïticos in Spain.

Since 1984, ISSP has grown to 41 nations, the founding four—Germany, the United States, Great Britain, and Australia—plus Austria, Ireland, Hungary, the Netherlands, Israel, Norway, the Philippines, New Zealand, Russia, Japan, Bulgaria, Canada, the Czech Republic, Slovenia, Poland, Sweden, Spain, Cyprus, France, Portugal, Slovakia, Latvia, Chile, Denmark, South Africa, Switzerland, Venezuela, Brazil, Flanders, Finland, Mexico, Taiwan, Korea, Uruguay, the Dominican Republic, Croatia, and Turkey. In addition, East Germany was added to the German sample upon reunification. Past participants not currently active include Bangladesh and Italy.

ISSP Themes

Since 1985, ISSP research has covered many key topics across a variety of disciplines.

The first theme in 1985 was on the role of government and covered (a) civil liberties, (b) education, (c) welfare and social equality, and (d) the economy.

The second theme in 1986 was on social networks and support system. It contained detailed behavioral reports on contacts with various friends and relatives and then a series of questions about where one would turn for help when faced with various problems.

The third module in 1987, on social equality, concerned beliefs about what factors affect one's chances for social mobility (e.g., parental status, education, race, etc.), explanations for inequality, assessments of social conflicts, and related questions.

The fourth module in 1988 covered the impact on the family of the changing labor force participation of women. It included attitudes on marriage and cohabitation, divorce, children, and child care and special demographics on labor-force status, child care, and earnings of husband and wife.

The fifth module in 1989 on work orientations dealt with motivations to work, desired job characteristics, unemployment, job satisfaction, and working conditions.

The sixth module in 1990 repeated the role of government theme. By replicating substantial parts of earlier modules (approximately two-thirds), ISSP not only has a cross-national perspective but also a longitudinal perspective. One is able not only to compare nations and test whether similar social science models operate across societies but also to learn if there are similar international trends and whether parallel models of social change operate across nations.

The seventh module in 1991 covered the impact of religious beliefs and behaviors on social, political, and moral attitudes.

The eighth module in 1992 replicated and extended the 1987 social equality module.

The ninth module in 1993 was on the environment. It included an environmental knowledge scale plus attitudinal and behavioral measures.

The 10th module in 1994 repeated the 1988 module on gender, family, and work.

The 11th module in 1995 was on national identity. It assessed nationalism and patriotism, localism and globalism, and diversity and immigration.

The 12th module in 1996 was the second replication of the role of government.

The 13th module in 1997 was the first replication of the 1989 module on work orientations.

The 14th module in 1998 was the first replication of the 1991 religion module.

The 15th module in 1999 was the second replication of the social inequality module fielded in 1987 and 1992.

The 16th module in 2000 was the first replication of the 1993 environment module.

The 17th module in 2001 was related to, but not a strict replication of, the 1986 social relations and social support module.

In 2002 the 18th module was the third replication of the gender, family, and work module.

In 2003 the 19th module was the first replication of the 1995 national identity module.

In 2004 the 20th module was a new module on citizenship.

In 2005 the 21st module was the second replication of the work orientation module,

In 2006 the 22nd module was the 3rd replication of the role of government module.

In 2007 the 23rd module was a new module on leisure activities.

In 2008 the 24th module was the second replication of the religion module.

ISSP marks several new departures in the area of cross-national research. First, the collaboration between organizations is not special or intermittent, but routine and continual. Second, while necessarily more circumscribed than collaboration dedicated solely to cross-national research on a single topic, ISSP makes cross-national research a basic part of the national research agenda of each participating country. Third, by combining a cross-time with a cross-national perspective, two powerful research designs are being used to study societal processes.

More than 2,300 publications based on the ISSP are listed in a bibliography available at http://www.issp.org/public.shtml.

Tom W. Smith

Further Readings

International Social Survey Programme: http://www.issp.org

Zentralarchiv für Empirische Sozialforschung, University of Cologne: http://www.gesis.org/ZA/index.htm

Internet Pop-Up Polls

Internet pop-up polls receive their name from their inherent function. These polls appear on the user's screen in a new browser window, which is triggered by accessing the content of a designated Web page. Like all Internet surveys, these surveys can reduce the time and cost of conducting research by streamlining the data collection process. Utilizing a dynamic programming language such as PHP (Hypertext Preprocessor) or PERL, the researcher can create a Common Gateway Interface (CGI) script that either emails the survey responses back to the administrator or automatically drops the entries into a pre-scripted database.

Pop-up polls can improve response rates by creating more attractive and easy-to-use forms. Specifically, they allow participants to view the base Web page while simultaneously taking an associated survey. However, studies have shown that people tend to respond more to relatively plain Internet surveys that load quickly rather than more elaborate surveys that necessitate longer load times.

Some researchers directly code their surveys in HTML (Hypertextual Markup Language). HTML editors often are easier to use and achieve the same effect. Microsoft's FrontPage or Mozilla's SeaMonkey both provide a point-and-click interface that allows researchers to easily develop questionnaires and the necessary CGI scripts. Various online tutorials are available to develop either independent or dependent HTML coding abilities (see Further Readings at the end of this entry). However, in practice, simple trial and error is usually the best training method. This training is especially useful when dealing with the many technical hurdles one encounters with Internet pop-up polls.

If the HTML pop-up poll is viewed as a protagonist, the antagonist would be the pop-up blocker. This is a formidable adversary that prevents most pop-up surveys from being viewed. To achieve this, the pop-up blocker does one of two things: Either it embeds code into the HTML script preventing additional windows from opening, or it alters the action code, which in HTML is the *window.open()* command, thus preventing the survey from ever popping up. Developers avoid this problem by either using DHTML or JavaScript to create floating banners that avoid most pop-up blockers by adding additional layers rather than windows to the base Web page. This reclassification can work in tandem with proper user warnings to ensure that most pop-up surveys are properly viewed. Once this technical obstacle is overcome, one must then address the inherent social obstacles to meaningful Internet survey research.

A major problem with Internet pop-up polls is that Internet users are not representative of the broader population, and heavy Internet users are not representative of lighter Internet users. Thus, coverage bias is a concern with Internet research, in general, and specifically with HTML pop-up polls because they require a certain level of familiarity with the Web interface to successfully interact with the survey instrument. Additionally, the lack of direct oversight means that potential abuse by respondents (e.g., responding multiple times to the same survey) is more likely in a Web-based environment.

Even given these limitations, HTML pop-up polls do have valid and legitimate uses. For example, pop-up polls can be an effective way to survey traffic to a given Web site. Even though the technology sometimes appears to be overwhelming, it can be harnessed to create unique survey instruments that can achieve ends that are beyond traditional means. The Internet is here to stay and the HTML pop-up poll is a likely bridge between what Internet research was and what it will become.

Bryce J. Dietrich

See also Coverage Error; Email Survey; HTML Boxes; Internet Surveys; Sampling Bias; Self-Selected Sample; Survey

Further Readings

O'Sullivan, M. (2002, January). *HTML popup windows*. Retrieved February 28, 2007, from http://www.htmlcodetutorial.com/linking/linking_famsupp_70.html

Schonlau, M. (2002). *Conducting research surveys via e-mail and the Web*. Santa Monica, CA: RAND.

Sue, V., & Ritter, L. A. (2007). *Conducting online surveys*. Thousand Oaks, CA: Sage.

INTERNET SURVEYS

Internet surveys refer to surveys that sample respondents via the Internet, gather data from respondents via the Internet, or both. Using the Internet to conduct survey research provides a great many opportunities and a great many challenges to researchers.

Background and Overview

Sample surveys have developed considerably over the past 70 years and have become the major source for the vast majority of empirical data, available today, on society, opinions, economics, and consumer preferences. Until the 1970s almost all survey work was carried out by pencil-and-paper questionnaires. Most of the collection was by means of face-to-face personal interview visits at the respondents' home or business. A small part of survey collection was by self-administered questionnaires, sometimes delivered and collected by interviewers and sometimes collected via mail. In recent times electronic telecommunications have become a predominant factor in practically all aspects of modern life, especially since the beginning of the 21st century. Sample surveys are no exception, and the widespread use of the telephone as a prime mode of communication, for at least the past 40 years, has had an important influence on survey practice. In fact, the telephone survey has become the major mode of collection in the sample survey field, especially in North America and Western Europe, both for surveys of households and individuals and for surveys of establishments. Other modes of advanced telecommunication, such as Internet, email, videophone, fax, and mobile phones, are fast becoming important supplements and even competitors to the fixed line telephone.

Internet surveys, sometimes termed *Web surveys* or *WWW surveys*, have fast become an important component of the trend to replace face-to-face interviewing, as the main mode of collection in survey work, with *telesurveys*—surveys carried out by modern telecommunication methods. The growing widespread access to the Internet and its extended use for a variety of purposes, particularly in Western Europe and North America, together with the rapid technological development of advanced browsers, XML, ADSL, and Java technology have brought about the continually increasing massive use of the Internet for survey work. The Internet survey can be viewed as a considerably enhanced replacement of the email survey, where text questionnaires are emailed to respondents, who are then asked to return the completed questionnaire by email. However, the Internet survey overcomes many of the inherent limitations of email surveys. The possibilities of visual and audio stimulation, the online interactive capabilities, and the potential of enhanced skip patterns available in the design of an Internet survey make it an extremely powerful survey data collection tool, far superior to the email survey. On the other hand the Internet survey may often suffer from serious problems of coverage, representativeness, and nonresponse bias.

Advantages

The major advantage of the use of the Internet in data collection is the very low cost per respondent, as compared to other modes of data collection. This has made the Internet survey an extremely attractive option to a wide range of survey researchers, primarily in the areas of opinion polling and market research, where the principles of probability sampling are not always considered as being of prime importance and large numbers of respondents are judged as valuable. The initial set-up costs entailed in the design of high-quality collection instruments via the Internet may be somewhat higher than those required for the design of paper questionnaires or computer-assisted telephone interviewing (CATI) instruments. However, this is more than offset by the current operational savings, due to self-administration of the survey instrument. The savings in the direct costs of interviewers, their training and control, are substantial. While other self-administered instruments, such as mail questionnaires and simple email collection, share with Internet surveys the advantage of not requiring the intermediary function of interviewers, for Internet surveys the costs involved in the control of unit and item nonresponse, callbacks, and editing are minimal and lower, in general, than those for other self-administered modes.

An important advantage of the use of the Internet for data collection lies in the advanced enhancements of the visual and aural aspects of the collection instrument. The use of color, animation, and even video and audio effects can, if used with care, facilitate the completion of the questionnaire for the respondent. The real-time interaction between the collection instrument and the respondent is a definite improvement over the fixed form of questionnaire required by other modes of data collection. Thus the use of drop-down menus, the possibilities to refer easily by hyperlinks and radio boxes or buttons to instructions and classifications, the possibilities to display photos of products or magazine covers, and other features all make the task of completing an Internet questionnaire much easier than for conventional instruments. Online editing, logical checks, and complex skip patterns can be employed in ways that are virtually invisible to the respondent.

However, some words of warning are in order with respect to the possibility of overkill in the use of these enhancements. The design of a good Internet questionnaire is a difficult and complex task, requiring a combination of survey research experience and advanced technical knowledge, far beyond that required for the design of conventional survey instruments. Also it is important to recognize that there are important differences in the design principles of the visual outlay for traditional questionnaires and the features of screen design. It is all too easy to cram bewildering amounts of information, complex instructions, and too many response alternatives within a single overcrowded screen, resulting in frustration for the respondents and high rates of break-offs (partial completions), missing data (item-nonresponse), or both. Finally it should be noted that variations in users' screen size and resolution, in operating systems, and in browsers may complicate the design and require simpler options.

Other advantages of Internet surveys are in the direct processing of the collected data in electronic form, bypassing the tedious and error-prone processes of data-capture, editing, coding, and logical checks required in traditional data collection methods. Some of these advantages are obtained to a certain degree by other modern computer-assisted collection methods (e.g., CATI and computer-assisted personal interviewing [CAPI]), though the Internet is usually associated with a more efficient method of data processing.

Finally, although not always perceived as such by the respondents, the Internet does provide for advanced methods of ensuring confidentiality, far beyond those afforded by conventional collection methods. The use of personal identification numbers or passwords provided to pre-designated respondents, secured access, and other technical enhancements can ensure that the individual data collected are not accessible to anyone except to the data collector. The widespread use of the Internet in nonsurvey contexts, for e-commerce purchases, registration, and other commercial activities, in which often full credit card details are required, with only a relatively minor extent of electronic fraud, should convince even the most skeptical that the data they provide for Internet surveys can be very secure.

Problems and Challenges

The main drawbacks in the use of the Internet for survey work are problems of coverage, lack of suitable sample frameworks, and nonresponse. The problems of coverage and coverage error are definitely the most serious of these problems for the vast majority of Internet surveys. While there has been a spectacular increase in the use of the Internet in the past decade, access and use of the Internet is still far from universal among households, even in well-developed countries. Accurate and up-to-date estimates of Internet use are hard to come by, but even the highest current estimates of Internet use (at least once a month) by households in North America in 2007 do not exceed 70% and fall below 60% for most countries in Western Europe. Numerous empirical studies have shown that, besides the fact that Internet coverage is relatively low, it is highly differential with respect to many socioeconomic variables and also with respect to attitudes and opinions. Thus, Internet users differ considerably from nonusers, in many important ways. This implies that the extent and nature of Internet undercoverage makes the use of the Internet completely inadequate as a tool for obtaining data on which valid inferences can be made, in a large number of cases, especially for opinion polling and attitude surveys, and in any survey that is measuring a construct related to education and income. Even in the cases mentioned next, when coverage is not an issue, it will often be difficult to select a probability sample, because of the lack of a complete adequate sampling frame.

It should be noted that the term *Internet survey* covers a wide range of different possibilities to use the Internet as a data collection tool in sample surveys, which may differ with respect to the potential impact of coverage and frame problems. Thus in many cases, the Internet may be used as an alternative mode of collection, within a mixed-mode collection process. Respondents may be given the option of responding via the Internet, at the time of their choice, rather than by CATI, by touchtone data entry or by a self-administered mail questionnaire. Although the mixed-mode survey may combine some of the advantages of the Internet mode of collection, such as easing the respondent's task, with the advantages of non-Internet surveys, such as enhanced representativeness and availability of a good sampling frame, it will, in general, be more expensive and entail a complex design.

When the Internet is used as the unique mode of collection, there may be significant differences in the way the target population is defined, with implications for the problem of undercoverage. In some cases the survey relates to a well-defined population, such as employees of a business, members of an organization, or students of an educational establishment. In many of these cases adequate sampling frames will be available, such as institutional listings of email addresses of employees or students. In this case, probability sampling can be used, and providing nonresponse is well controlled, scientifically valid inferences can be obtained from an Internet survey. Another type of survey in which the Internet can be used efficiently with few problems of coverage is that of surveys of businesses, establishments, or organizations. The Internet coverage of institutions, businesses, and organizations is virtually complete, with the possible exception of very small businesses. While frames of businesses and institutions are generally available, they will often not include email addresses, so an initial approach by mail or telephone may be required.

Another case in which coverage problems are nonexistent is that of surveys in which the population is defined a priori as that of Internet users, such as surveys of customer satisfaction or Internet site evaluation. Here the problem of the sampling frame may be severe if the survey is of all users of the Internet, since there are no general lists of Internet users. Internet service providers will obviously provide lists of their customers only for surveys commissioned by them. If the population survey is defined as those accessing a specific site or using a specified Internet service, the frame problem can be solved by sampling systematically from the users at the time of entry, with multiple selections avoided by the use of cookies. For these types of surveys the problems of nonresponse are paramount, and typically very low levels of response are obtained.

However, in many cases the Internet is used for survey work in a relatively uncontrolled way, without rigorous definitions of a survey population and therefore without a scientifically designed sample method. This has serious implications for the possibility to make valid well-based inferences about the subjects studied. At the extreme, the ubiquitous Internet versions of entertainment polls, customer preference and feedback surveys, and unrestricted self-selection and volunteer opt-in opinion polls are set up with appeals to anyone who is willing to do so to respond. No valid inference can be obtained from such attempts, which do not bear an association with scientific surveys. Unfortunately, a large number of well-publicized appeals to respondents to respond to these so-called Internet surveys have produced very high numbers of respondents, who cannot, however, be considered as representative of any well-defined population. The fact that the number of respondents thus obtained is very large (sometimes in the hundreds of thousands) is represented as sufficient to ensure reliability of the results, whereas, in fact there is not even a possibility to evaluate the biases inherent in such efforts. The attempts to overcome the lack of any scientific sampling design by post-stratification or other methods of weighting or imputation are not effective to ensure the correction of these biases.

Future Developments

The Internet is a powerful and inexpensive method of data collection, with many advantages and enormous potential in cases where it can be used in the context of probability sampling. However, in many current applications, coverage and frame problems prevent its being used for probability sampling–based surveys to ensure valid inferences. Future developments may change the situation. Thus the proposed introduction of a unique universal personal communications number for use with all modes of telecommunication (fixed-line and mobile phones, fax, and email) may solve many of the problems associated with coverage and absence of frameworks, at least for multi-mode surveys.

Gad Nathan

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Coverage; Coverage Error; Drop-Down Menus; Email Survey; Face-to-Face Interviewing; Missing Data; Mixed-Mode; Multi-Mode Surveys; Partial Completion; Propensity-Weighted Web Survey; Radio Buttons; Sampling Frame; Undercoverage; Web Survey

Further Readings

Clayton, R., & Werking, G. S. (1998). Business surveys of the future: The World Wide Web as data collection methodology. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nichols, et al. (Eds.), *Computer-assisted survey information collection* (pp. 543–562). New York: Wiley.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

Dillman, D. (2007). *Mail and Internet surveys: The total design method* (2nd ed.). New York: Wiley.

Nathan, G. (2001). Telesurvey methodologies for household surveys—a review and some thoughts for the future. *Survey Methodology 27*, 7–31.

Interpenetrated Design

An interpenetrated survey design is one that randomly assigns respondent cases to interviewers. This is done to lower the possibility that interviewer-related measurement error is of a nature and size that would bias the survey's findings. This type of design addresses survey errors associated with the survey instrument and the recording of responses by the interviewer. One way to reduce *subjective* interviewer error is to develop a survey using an interpenetrated design—that is, by ensuring a random assignment of respondents to interviewers. Surveys employing an interpenetrated design, when such is warranted, will tend to reduce the severity of interpretation errors resulting from the conflation of interviewer bias with some other statistically relevant variable that might serve as a basis for assigning respondents. It will also typically reduce the overall standard error of response variance, especially for types of questions that inherently require some judgment or interpretation in recording by the interviewer.

Example of an Interpenetrated Survey Design

Assume a survey of 100 women from known high-risk populations (e.g., low income, substandard education, history of domestic violence), who are being queried about their tobacco use. The survey will be administered face-to-face by five interviewers and will feature a mix of demographic and binary-response questions, as well as several open-ended questions about the respondents' psychosocial triggers for smoking that the interviewer will interpret and assign a clinical risk index score.

In an interpenetrated design, the 100 women will be randomly assigned to each of the five interviewers. This means that any potential skewing of recorded results arising from bias or judgment by any single interviewer will be relatively equally shared by all of the respondents assigned to that interviewer and could therefore be considered "background noise" in terms of finding correlations within and among classes in the data. By contrast, in a noninterpenetrated design, it is possible that a correlating variable or class could be overemphasized or underemphasized by the relative weight of interviewer bias across a nonrandom assignment of respondents to interviewers.

For example, if all pregnant women queried about their tobacco use were assigned to a single female nurse interviewer who believes smoking is a social vice and not a chemical addiction, the nurse-interviewer's own subjective bias might contribute to Type I or Type II error for the class of pregnant women, relative to the survey's working hypothesis, or the bias might introduce systemic error into the response-variation rate for the class of pregnant women assigned to that interviewer. An interpenetrated design, in this example, would decrease the likelihood that one interviewer's behavior will contribute in a statistically significant way to analytic error.

Challenges in Implementing an Interpenetrated Design

The use of an interpenetrated design can mitigate the inflation of *statistical* error engendered from *systematic* error in survey design, for surveys with measurement tools or questions that fail to adequately control for interviewer bias, in cases where such bias could affect findings.

It can be difficult to engineer an effective interpenetrated design, however. There may be situations, particularly with large-scale face-to-face surveys, when geography or interviewer expertise with a particular class of respondent reduces the design's capacity to fully randomize the assignment of interviewer to respondent. There may be some benefit to determining whether a mixed strategy might be appropriate, with a partial randomization along respondent demographic, location, or cohort lines that are not believed to be relevant to the hypothesis of the survey or in its final analysis. As with any survey design, the question of which variables should be considered relevant must be approached with great caution.

Colm O'Muircheartaigh and Pamela Campanelli, in their targeted meta-analysis of the British Household Surveys of the 1990s, concluded that there was a significant increase in the inflation of variance rooted in measurable interviewer effects that were comparable in scope to the variance attributed to survey-design effects. Their findings, and the work of other statisticians, suggest that a clear understanding of the interviewer effect on the intraclass correlation (rho) is necessary for effective survey designs, and the use of an interpenetrated design can be quite effective in mitigating this interviewer effect at a statistical level.

However, depending on the nature of the survey, it can become prohibitively expensive or complex to employ an interpenetrated design. Surveys requiring face-to-face interviewing, for example, might generate substantial interviewer travel costs if the respondents are located across a broad geographic territory. Even in telephone interviews, the additional resources needed to ensure appropriate randomization with small groups of interviewers—perhaps through a stratified sampling model—can increase the burden of developing the survey and may generate additional analysis to ensure that the interpenetrated design contributed in a statistically significant way to an improvement in precision for the survey's final analysis. With a telephone survey, it is hardly ever feasible to control the sample so that each interviewer works only a randomly assigned subset of cases.

Given the increase in costs and complexity that can occur when implementing an interpenetrated design, researchers should determine carefully—using appropriate statistical tools—whether the very real benefits of an interpenetrated design outweigh the added costs and complexities of developing and implementing one.

Jason E. Gillikin

See also Bias; Correlation; Covariance; Design Effects (*deff*); Face-to-Face Interviewing; Interviewer Effects; Interviewer-Related Error; Noncausal Covariation; Open-Ended Question; Questionnaire-Related Error; Random Assignment; Research Hypothesis; Respondent–Interviewer Rapport; ρ (Rho); Standard Error; Type I Error; Type II Error

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

O'Muircheartaigh, C., & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(3), 437–446.

Stokes, L. (1988). Estimation of interviewer effects for categorical items in a random digit dial telephone survey. *Journal of the American Statistical Association*, *83*(403), 623–630.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

INTERRATER RELIABILITY

The concept of interrater reliability essentially refers to the relative consistency of the judgments that are made of the same stimulus by two or more raters. In survey research, interrater reliability relates to observations that in-person interviewers may make when they gather observational data about a respondent, a household, or a neighborhood in order to supplement the data gathered via a questionnaire. Interrater reliability also applies to judgments an interviewer may make about the respondent after the interview is completed, such as recording on a 0 to 10 scale how interested the respondent appeared to be in the survey. Another example of where interrater reliability applies to survey research occurs whenever a researcher has interviewers complete a refusal report form immediately after a refusal takes place and how reliable are the data that the interviewer records on the refusal report form. The concept also applies to the reliability of the coding decisions that are made by coders when they are turning open-ended responses into quantitative scores during open-ended coding.

Interrater reliability is rarely quantified in these survey examples because of the time and cost it would take to generate the necessary data, but if it were measured, it would require that a group of interviewers or coders all rate the same stimulus or set of stimuli. Instead, interrater reliability in applied survey research is more like an ideal that prudent researchers strive to achieve whenever data are being generated by interviewers or coders.

An important factor that affects the reliability of ratings made by a group of raters is the quantity and the quality of the training they receive. Their reliability can also be impacted by the extent to which they are monitored by supervisory personnel and the quality of such monitoring.

A common method for statistically quantifying the extent of agreement between raters is the intraclass correlation coefficient, also known as Rho. In all of the examples mentioned above, if rating data are not reliable, that is, if the raters are not consistent in the ratings they assign, then the value of the data to researchers may well be nil.

Paul J. Lavrakas

See also Coding; Content Analysis; Open-Ended Question; Refusal Report Form (RRF); ρ (Rho)

Further Readings

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

INTERVAL ESTIMATE

Interval estimates aim at estimating a parameter using a range of values rather than a single number. For example, the proportion of people who voted for a particular candidate is estimated to be 43% with a margin of error of three (3.0) percentage points based on a political poll. From this information, an interval estimate for the true proportion of voters who favored the candidate would then consist of all the values ranging from a low of 40% to a high of 46%—which is usually presented as (0.40, 0.46). If the interval estimate is derived using the probability distribution of the point estimate, then the interval estimate is often referred to as a "confidence interval" where the "confidence coefficient" quantifies the probability that the process and subsequent derivation will produce an interval estimate that correctly contains the true value of the parameter.

While point estimates use information contained in a sample to compute a single numeric quantity to estimate a population parameter, they do not incorporate the variation in the population. Interval estimates, on the other hand, make use of the point estimate along with estimates of the variability in the population to derive a range of plausible values for the population parameter. The width of such intervals is often a function of the "margin of error," which is itself a function of the degree of confidence, the overall sample size, and sampling design as well as the variability within the population. In practice, intervals that are narrower usually provide more specific and useful information about the location of the population parameter as compared to wider intervals that are often less informative or more generic (e.g., the population proportion of voters in favor of a candidate is between 0 and 1 would be an interval estimate that is not informative).

Interval estimates can be derived for any population parameter, including proportions, means, totals, quantiles, variances, regression parameters, and so on. The generic format of an interval estimate for the population parameter θ can be written as $\hat{\theta} \pm DV \times SE(\hat{\theta})$ where DV represents a distribution value determined by the sampling distribution of the estimator $\hat{\theta}$, and *SE* refers to the standard error of the estimator. Many interval estimates are in fact symmetric around the corresponding point estimate (i.e., as is generally true for means, totals, and proportions based on simple or stratified random sampling designs), but this property is not universal. For example, if interest is given to estimating the variability in birth weights in a population using a simple random sample of hospital birth records, then estimates of σ^2 will be based on a chi-squared distribution and thus will generally be asymmetric about the point estimate.

As example, suppose that interest is given in estimating the average household 6-month out-of-pocket dental expenses within a midwestern U.S. state. An interval estimate is to be derived using a sample of 10 households that are randomly selected from each of eight geographically defined strata. Sample means, variances, and stratum sizes (i.e., numbers of households) are provided in Table 1.

A $(1-\alpha) * 100\%$ confidence interval estimate for the mean household dental expenses is given by $\bar{y} \pm t_d (1-\alpha/2) \times [v\hat{a}r(\bar{y})]^{1/2}$ where the degrees of freedom d are computed as the number of sampled units (i.e., 80) minus the number of Strata (i.e., 8). From the information in Table 1, a 90% confidence interval estimate of the mean 6-month household out-of-pocket dental expenses based on this stratified sample of 80 homes from 8 strata is $\$109.83 \pm 1.667 * (13.26)^{1/2}$ or equivalently, (\$103.76, \$115.90), implying that one can be 90% confident that the true mean 6-month household out-of-pocket dental expenses is between approximately \$104 and \$116.

Table 1 Data from a stratified random sample of 10 households in each of 8 geographic strata to be used to estimate the mean household 6-month out-of-pocket dental expenses

Stratum	Stratum Size (Households)	Sample Size	Sample Mean	Sample Variance
1	150,000	10	\$88.50	660.49
2	100,000	10	\$108.90	900.00
3	50,000	10	\$110.25	576.00
4	50,000	10	\$100.00	784.00
5	50,000	10	\$106.75	729.00
6	75,000	10	\$176.40	1,296.00
7	25,000	10	\$200.20	1,444.00
8	300,000	10	\$98.70	529.00
	Sample variance of the sample mean:			**13.26**
	Stratified sample mean estimate:			**\$109.83**

Trent D. Buskirk

See also Confidence Interval; Margin of Error; Model-Based Estimation; *p*-Value; Point Estimate; Population Parameter; Standard Error; Variance

Further Readings

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole.

Interval Measure

An interval measure is one where the distance between the attributes, or response options, has an actual meaning and is of an equal interval. Differences in the values represent differences in the attribute. For example, the difference between 3 and 4 is the same as the difference between 234 and 235. Interval measures have fixed measurement units, but they do not have a fixed, or absolute, zero point. Because of this, it is technically not correct to declare that something is so many times larger or smaller than something else, although this often is done nonetheless.

Unlike other less sophisticated levels of measurement (e.g., nominal and ordinal measures), interval measures have real meaning. The relationship between the value and attribute is meaningful. For instance, temperature (Fahrenheit or Celsius) can be divided into groups of one degree and assigned a different value for each of the intervals such that anything from 50 degrees to 50.99 degrees has a value of 50. The distance between 50 and 51 degrees has an actual value, one degree. On the other hand, one cannot say that 90 degrees is twice as hot as 45 degrees because there is not an absolute zero.

Within social science research, interval measures are not particularly common because there are only a limited number of attributes that can take on an interval form. When used, they tend to be based on constructed measures like intelligence tests or standardized tests. Another common interval measure is a year. Some within the behavioral sciences use the Rasch model to create interval measures from count data.

It is often common in the research literature to see ordinal measures treated like interval measures. Many ordinal measures have characteristics that suggest that they could be thought of as interval measures. Many subjective rating scales are treated as interval measures. For instance, it is common to see measures using a 7- or 5-point scale, like the "strongly agree," "agree," "disagree" Likert scale battery interpreted as interval. It is easy to see why the interpretation exists, but it is not possible to conclusively state that the different values are exactly equally spaced. That said, many argue that with sufficient testing, it is possible to gain a sense of confidence that the intervals on an ordinal scale are close enough to be treated as though they are equivalent.

Interval measures allow more transformations than nominal and ordinal measures and are generally considered stronger measures, thereby supporting the use of parametric statistical procedures. This means that interval measures must also satisfy the assumptions of nominal and ordinal measures. Therefore, the interval measures 3 and 4 can be thought of in terms of these other measures, such that 4 is considered to have more of the attribute being measured than 3 as is found in ordinal measures. At the same time, a nominal interpretation would suggest that all things assigned 3 would have similar attributes to one another. Because of this, interval measures can always be transformed into ordinal or nominal measures.

Interval measures allow fewer transformations than ratio measures and are considered weaker measures. The central tendency of an interval measure can be represented by its mode, median, or mean. Usually, the mean is considered to provide the most information. Interval measures can also be added and subtracted.

James W. Stoutenborough

See also Level of Measurement; Likert Scale; Mean; Median; Mode; Nominal Measure; Ordinal Measure; Ratio Measure

Further Readings

Babbie, E. R. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*, 65–72.

INTERVIEWER

Interviewers are survey staff who contact the people from whom the study seeks to gather data (i.e., respondents) to explain the study, encourage them to participate, and attempt to obtain a completed interview. Once a sample member agrees to participate in a survey, the interviewer is then responsible for administering the survey questions (i.e., survey instrument) to the respondent. At times, the skill sets necessary to successfully complete these two tasks—sample recruitment and data collection—differ in conflicting ways. In encouraging participation, interviewers must adapt (i.e., tailor) their approach to gain cooperation based on their interaction with the respondent, whereas for administering the questionnaire in most surveys they are encouraged to use a standardized, scripted approach.

Traditionally, there have been two types of survey interviewers: telephone interviewers and field interviewers. Telephone interviewers administer survey instruments over the telephone, whereas field interviewers go to a respondent's home (or business for business surveys) to contact respondents and complete the in-person interview face-to-face. More recently and at a growing pace, interviewers are also being used to provide technical assistance (e.g., help desk) to self-administered surveys, such as mail or Web.

This entry presents the responsibilities of interviewers, along with the various skills needed and the training and supervision of interviewers. Next, this entry discusses common interview techniques and the impact of interviewers on the data collected. Lastly, this entry addresses the importance of interviewers to the data collection effort.

Responsibilities and Skill Sets

Interviewers have multiple responsibilities on any survey effort. Their initial responsibility is to contact respondents, and when necessary, screen for the eligible respondent, and also to provide detailed information to help explain the survey. During this initial exchange, interviewers attempt to gain cooperation from respondents so that they will complete the survey task.

Interviewers are also responsible for converting cases where a respondent, or gatekeeper (someone who is keeping the respondent from talking directly with the respondent), has initially refused to participate. Another key responsibility for interviewers is to administer the survey instrument and to do so in an unbiased manner while correctly recording the responses obtained from the respondent into the survey instrument. The successful execution of these responsibilities contributes heavily to the success of the overall data collection effort.

Successful interviewers, regardless of mode of interview, need to possess a set of varied skills. Interviewers must be convincing, professional, friendly, knowledgeable, and empathetic. The interviewers' attitude can also have a large impact on their overall success. They must also possess persuasion skills and use these skills to initially gain cooperation, as well as to collect complete data and encourage continued participation as the interview progresses.

Interviewer Training and Supervision

Although the location and content of training for telephone and field interviewers can vary, the overall information conveyed about the survey and the techniques used to convey information are similar across survey administration modes. Telephone interviewers are usually located in a centralized facility and are trained on-site at a call center. However, field interviewers are often decentralized and are often brought to a centralized location, such as a hotel, for training.

The length of interviewer training often differs based on the mode of survey administration. Generally, telephone interviewer training is shorter than that for field interviewers because of the nature of the work. Telephone interviewers need to know how to use a computer and telephone, whereas field interviewers also have to know how to approach and gain access to residences or businesses, complete time and expense reports, as well as how to upload (or transmit) survey data to a centralized database. These differences in job responsibilities account for the differences in training content and length and in compensation levels. Both types of interviewers, however, must also undergo general training, which introduces them to the job and the field of survey research, before undergoing project-specific training.

However, the information conveyed about the survey effort and the training techniques used is often very similar. It is understandable that regardless of survey administration mode, all interviewers need to be trained on the basic purpose and importance of the survey, as well as in administration techniques. Further, all interviewers also need to have hands-on practice with the systems used to collect the data, such as computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) systems. Regardless of mode, trainers often use similar techniques to convey the information that interviewers need for their job. Trainer-led lecture, small group discussion, mock interviews that all interviewers participate in (also called "round robin" interviews), paired practice, verbal and written quizzes, and hands-on practice with systems are techniques used in many training sessions. There have been several studies that have looked at the impact of interviewer training on subsequent job performance. Some of this literature focuses on the impact of using various training techniques for refusal avoidance and conversion on subsequent performance.

While the performance and monitoring of interviewers are important regardless of mode, the techniques used to supervise telephone and field interviewers differ. Telephone interviewers are often considered easier to manage because they are centrally located and supervisors can interact with them during each interviewing shift. It is also possible to silently monitor telephone interviews, using both audio and visual displays, in a way that is undetectable by the interviewer and respondent. This can result in provision of feedback to interviewers soon after they are observed, which may help improve their performance. It is also possible to meet in person with telephone interviewers to provide both individual feedback and feedback on how the entire data collection is proceeding.

However, due to the decentralization of field interviewers, many of these same techniques are not available to supervise field interviewers. While telephone meetings can be held with multiple field interviewers simultaneously, it can be difficult to get interviewers engaged during such a forum. Further, to monitor the work of field interviewers, supervisors must actually physically join a field interviewer when they are performing their duties. (Unlike with telephone interviewing, the respondent is thus aware of the presence of the supervisor.) Other monitoring can be done via computer-assisted mechanisms, for example, where the computer used for interviewing automatically makes an audio recording for a section (or multiple sections) of the interview.

For both types of interviewers, however, timely and complete feedback on their performance is critical to the maintenance and improvement of their skills. Supervisors of both telephone and field interviewers should provide this information routinely to their interviewers and provide routine coaching and mentoring in an attempt to maintain and improve interviewer skills, thereby helping the overall data collection effort.

Common Interviewing Techniques

There are multiple techniques used to conduct an interview, and the type of technique used in a particular study is dependent upon the study's purpose. Interviews that are more qualitative in nature and whose purpose is to gather information to be used in construction of a future survey instrument are often loosely structured, and the interviewer has a lot of discretion in the types of questions asked as well as the wording of particular questions. These kinds of interviews can be *unstructured* (where only general topics to cover are provided), *semi-structured* (where some questions are listed but interviewers have freedom to add, delete, or alter questions), or *structured* (where the questions need to be asked as listed).

Alternatively, for quantitative survey data collection efforts, there are *standardized* and *conversational* interviewing techniques. Each technique is appropriate for a different type of survey effort.

Standardized interviewing techniques are viewed by many survey professionals as the preferred method because they enable all respondents to receive the same stimuli (question and response wording) in the same way. In theory, this technique minimizes the impact of the interviewer such that any differences observed can be attributed to differences among respondents, and not to interviewer behavior. A benefit of standardized interviewing is that the effect of the interviewer on the data collected is decreased. Much has been written detailing the techniques used to ensure interviewers administer the survey instrument in a standardized way. These instructions focus on things such as reading all survey text verbatim, using nondirective probes to obtain complete information, recording all respondent answers exactly as provided, and administering the survey in a neutral, nonjudgmental manner.

Conversational interviewing involves following the survey text as written most of the time, but allowing for not asking or confirming the answer categories that may have been provided in an earlier part of the interview by the respondent through their comments or response to the questions. It also allows interviewers to provide unscripted explanations and definitions to clarify questions. For example, when asked if she attended college, the respondent answers, "Yes, I went to Michigan State," and three questions later the interviewer is supposed to ask what college the respondent attended, the interviewer could confirm, "So, you went to Michigan State University for college, is that correct?" instead of reading the question exactly as written. The idea is that the interview is more a conversation between interviewer and respondent, and by allowing for use of confirmatory statements it shows that the interviewer is actually listening to the respondent and not just asking questions and not paying attention to the answers, or other information, provided. A downside to conversational interviewing is that it does not control the impact of the interviewer as closely as standardized interviewing does, and this is viewed as problematic by some researchers.

Impact of the Interviewer on the Data Collected

As mentioned earlier, interviewers can have an impact on the quality, completeness, and accuracy of the data collected. That is, interviewers can contribute to survey error (the difference between the true value of what the researcher is trying to measure and the data that are obtained). Nonsampling error is difficult to measure, but there are three main ways used to detect interviewer-related error: (1) directly observing the interviewers, (2) examining the answers interviewers obtain (i.e., examining the association between the interviewer and the respondents' answers), and (3) validating the data collected with an independent (external) source.

It is important to realize that interviewers can affect both the variance and bias of the obtained estimates. The variability in interviewer administration can impact the variance of estimates, whereas demographic characteristics and other interviewer behaviors can affect bias.

The impact of interviewers on data can take several forms. Interviewers who accept a "don't know" response from respondents without probing or giving the respondent additional time to think about their answer can have more missing data. Also, interviewers who do not provide a probe, when one is called for, can impact the accuracy of the data. For example, if a respondent is asked the degree to which they agree with a statement—

strongly agree, somewhat agree, somewhat disagree, or strongly disagree—and the respondent says "agree," if the interviewer does not probe to see whether the respondent means "strongly" or "somewhat" agree and just chooses a response, then the data may not accurately reflect the respondent's view. Or if an interviewer directs a respondent toward an answer choice, that action can impact the variance of the estimate.

In many ways, the interviewer has to "teach" the respondent how to be a "good" respondent by providing verbal feedback on the respondent's actions during the interview. For example, if a respondent says "don't know" to questions and the interviewer never probes, then the respondent sees that providing more precise answers to the questions is not viewed as important. Further, if an interviewer lets respondents veer off of the interview topic repeatedly without trying to bring them back on task, then the respondents learn that this is acceptable behavior.

Interviewers can also impact the number of respondents who agree to be interviewed. Good persuasion skills are necessary for interviewers. If interviewers do not possess these skills and do not learn them during training, then the response rate can suffer as more respondents decline to participate.

Research has also investigated whether interviewer demographic characteristics, personality characteristics, or social skills can impact the survey result. In particular, interviewer race, religion and ethnicity, social status, education, and age have all been investigated. Although a significant amount of research has been conducted over multiple decades to look at the impact of interviewer demographics on survey data, few consistent significant results have been found. The only significant results that have been observed are when the interviewer characteristics are directly related to the questions being asked (e.g., gender of interviewer and questions about gender roles and equity). Research has also been conducted looking at the interaction between characteristics of the respondent and the interviewer, and in general this research finds few interactions and those that are found are of a complex nature and not easily disentangled. The general consensus seems to be that there is no demographic basis for choosing an interviewer.

Data collection costs are usually the largest piece of any survey budget. As such, the work performed by interviewers has a direct impact on the survey budget. If interviewers take more time to complete an interview than initially budgeted, then the overall project budget is likely to suffer unless other measures are taken to counter that overrun. Similarly, if interviewers complete interviews much more quickly than budgeted, this will have a positive effect on the project budget. Further, if interviewers are able to complete more interviews than initially projected, the need for additional interviewers and interviewing time may be reduced, which would also have a positive effect on the budget.

Importance of Interviewers to the Data Collection Effort

As the main visible representatives of the survey, and the people who are actually collecting the data, interviewers play a crucial part in data collection efforts. Interviewers are often the only contact that a respondent has with the survey team and as such need to provide a positive, professional image of the study. Interviewers are also a key part of obtaining an acceptable response rate by convincing respondents to participate in the survey and by conducting refusal conversions on respondents who initially refuse to participate. If an interviewer is disinterested or ineffective in obtaining cooperation from respondents, this can result in significant refusals or passive noncooperation, which can be detrimental to a study's response rate and data quality.

Lisa Carley-Baxter

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Conversational Interviewing; Gatekeeper; Interviewer Characteristics; Interviewer Effects; Interviewer Monitoring; Interviewer-Related Error; Interviewer Training; Interviewer Variance; Nondirective Probing; Refusal Avoidance; Refusal Conversion; Respondent; Respondent–Interviewer Rapport; Standardized Survey Interviewing; Survey Costs; Tailoring; Verbatim Responses

Further Readings

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.

Interviewer Characteristics

Interviewer characteristics refer to the personal attributes of the interviewer who is conducting a survey with a respondent. These attributes may include physical attributes, such as gender, age, and voice qualities, and attitudinal or behavioral attributes, such as confidence or friendliness. Both visual and audio cues are available to respondents in face-to-face surveys, but respondents have only audio cues in telephone surveys. The characteristics of the interviewer introduce additional factors into the interaction between interviewer and respondent that may affect data collection and data quality. Research has shown that interviewer characteristics affect unit nonresponse, item nonresponse, and response quality.

Physical Characteristics

Physical characteristics of the interviewer, such as age, gender, or race, may be used by the respondent in the decision whether to agree to the survey request and to set expectations of the interview experience. Studies on the effects of these attributes on interview outcomes show mixed results. There is evidence that older interviewers are more likely to be consistent in administering surveys and introduce less response variation. No consistent effects of gender have been found on data quality although female interviewers, on average, achieve higher response rates. There has been considerable study of interviewer race effects. The matching of characteristics of the interviewer to the respondent has been shown to improve respondent cooperation and data quality. Respondents appear to be more comfortable and thus cooperative with someone similar to themselves, especially in interviews on sensitive topics such as inequality and racial discrimination.

In telephone interviews, interviewer characteristics can only be conveyed through the audio interaction with the respondent. Physical characteristics that can be perceived over the phone include gender, age, and possibly, race and ethnic origin, as well as voice characteristics such as loudness and rate of speech. These characteristics can be measured both acoustically and through subjective perception. Acoustic measures of voice properties that have been studied include fundamental frequency of the voice sound waves, the variation in fundamental frequency, and measures of rate of speech and inflection. While acoustic voice measures have been found to help explain interviewer success, the measures studied have not been found to explain outcome of a particular contact. As with face-to-face interviewing, female interviewers tend to have higher response rates, but there is limited empirical evidence of gender effects on data quality. Race and ethnic origin of the interviewer may be conveyed through accents or articulation, but there is little research on these possible effects.

Attitudinal and Behavioral Characteristics

Attitudinal and behavioral characteristics of the interviewer have also been found to be related to survey response and cooperation. These characteristics are more difficult to measure, but they have been found to be significant predictors of response, as they represent the attributes of the interviewer that lead to establishing successful interviewer–respondent rapport. Attitudinal characteristics that have been studied include interviewer confidence, optimism, and persuasiveness. Behavioral characteristics have included attention to details, social skills, and interview behaviors such as points of engagement in survey introductions. Interviewer confidence, as measured either by the interviewer or by an observer, has been found to be positively related to survey success, although evidence is not conclusive. The effects of interviewer optimism, friendliness, agreeableness suggest that being overly open with a respondent leads to greater response variation. Interviewers who are socially skilled, without being overly sociable, are more likely to be able to tailor the interaction with the respondent and maintain data quality across interviews. A study that analyzed the effects of interviewer physical attributes, interviewer attitudes, and interviewer behaviors across multiple surveys and across countries found all three categories of characteristics to significantly predict interviewer response rates, with interviewer attitudes being the best predictor.

Attitudinal and behavioral characteristics can be conveyed over the phone, but they lack visual cues such as interviewer demeanor or expression. An interviewer may be judged as being friendly, conversational, or competent. Studies to identify the voice qualities of the best interviewers have included human ratings of voices by trained raters. Ratings have included both physical properties, such as pitch, rate of speaking, and loudness, as well as perceived attitudes, including pleasant to listen to, conversational, confident, competent, friendly, and enthusiastic. Early studies found lower refusal rates for interviewers who spoke rapidly, loudly, with standard American pronunciation, and who were judged as sounding confident and competent. A more recent study of survey introductions found listener ratings of greater confidence and voice breathiness to predict lower response rates, while variation across contacts, on multiple voice measures, was characteristic of high response rate interviewers.

Many studies have included interviewer experience as a characteristic to predict survey outcomes. Rather than a monotonic relationship of experience with higher response rates and data quality, there is indication that interviewers of middle-range experience are the most successful. Experience, as an interviewer characteristic, is clearly confounded with the physical, attitudinal, and behavioral characteristics discussed. The right combination of these characteristics is more likely to lead to success and experience on the job.

Effects

Interviewer characteristics contribute to defining the social interaction of the interviewer and respondent. Interviewer variation exists and contributes to survey outcomes. The mechanisms that define the effect of the characteristics are dependent on other survey conditions, including the respondent and the survey design.

Barbara C. O'Hare

See also Interviewer Effects; Interviewer-Related Error; Interviewer Variance; Respondent–Interviewer Rapport

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R. M., O'Hare, B. C., Gould-Smith, D., Benki, J., & Maher, P. (2008). Telephone interviewer voice characteristics and the survey participation decision. In J. Lepkowski, C. Tucker, M. Brick, E. De Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 385–400). New York: Wiley.

Hox, J. J., & de Leeuw, E. D. (2002) The influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 103–120). New York: Wiley.

Link, M. W., & Armsby, P. P. (2001, May). *Predicting persistence and performance among newly recruited telephone interviewers*. Paper presented at the annual conference of the American Association of Public Opinion Research, Montreal, Canada.

Shuttles, C. D., Skyrme, P., Lavrakas, P. J., Vallar, K., Haskins, C., & Wilkinson, D. (2006, January). *Hiring "The Right Stuff": Development of an assessment system for hiring effective interviewers*. Paper presented at the 2006 International Conference on Telephone Survey Methods II, Miami, FL.

Interviewer Debriefing

Interviewer debriefing is a process used to gather feedback from telephone and field (in-person) interviewers regarding a particular survey effort. As the project staff members who most closely interact with respondents, interviewers provide a unique perspective on how questions are answered by respondents and which questions may be difficult to ask or answer. They also can provide other, general observations about the administration of the survey instrument. These debriefing sessions can be held either in-person (as is usually the case for telephone interviewer debriefing sessions) or over the phone (as is often the case for field interviewer debriefing sessions).

Prior to conducting an interviewer debriefing, a member of the project staff—either the project director or the person who has managed the data collection effort—usually creates a detailed questionnaire for interviewers to complete prior to the session. Most of the questions should be short-answer (closed-ended) but also should include space so that the interviewer can provide feedback and concrete examples from their own experience in administering the survey on specified topics. Providing an additional open-ended question at the end of the questionnaire can also encourage the interviewer to comment on any other circumstances not covered in the debriefing questionnaire.

The debriefing questionnaire is given to interviewers by their supervisor, and they are instructed to complete it prior to the debriefing session. Interviewers are usually given several days to complete the questionnaire and are encouraged to initially spend a concentrated amount of time filling out the questionnaire and then continue adding to it in subsequent days as additional examples or comments occur to them. This process allows interviewers to spend multiple days thinking about their comments and allows them to relate those comments that are important to them. Further, quiet interviewers may be overlooked or not speak up much during the debriefing discussion, and gathering written comments from all interviewers allows project staff to get the entire range of feedback and not just feedback from the vocal interviewers.

A few days after handing out the debriefing questionnaire, project staff members meet with the interviewers to lead a discussion about the interviewers' experiences with respondents and the questionnaire, as well as to gather their comments on the data collection effort. This discussion is often organized like a focus group where the moderator attempts to get all participants involved in the discussion and where the discussion is directed by notes the moderator has prepared ahead of time. The information gathered during the interviewer debriefing session is only as good as the questions designed to elicit the information as well as the ability of the session leader or moderator to encourage participation and keep the discussion productive throughout the session.

For interviewer debriefing sessions, the questions on the debriefing questionnaire usually should form the structure of the discussion. It is important, however, for the project staff who are moderating or leading these discussions to be able to readily identify comments that need additional detail (or probing) as well as to readily identify comments that are off track so the discussion can be kept focused on the topic at hand. The written debriefing questionnaires are then collected at the end of the session.

Interviewer debriefing sessions can be audio- or videotaped for follow-up and archive purposes; however, in practice it is more likely that one or two project staff members will be assigned to take notes during the discussion. These notes, along with the written questionnaires, form the basis for the summary of the interviewer debriefing that can then be utilized by project management and the client or survey sponsor as part of the survey evaluation process. A key skill required for the project staff members assigned to summarize the interviewer debriefing comments is the ability to evaluate the comments in terms of what reasonably can and should be changed and what cannot or should not, as well as what issues are important for the overall survey administration and which are not. This sifting and evaluating of interviewer debriefing comments must be completed by a senior project staff member who understands the entire survey process and survey objectives in a way that interviewers, with their limited role on the project, usually cannot.

Interviewer debriefing sessions can also be opened up to include supervisor and monitors (for telephone interviews) or field observers (for field interviews). Similar to interviewers, these supervisory staff members offer a unique perspective on the data collection process. By virtue of their job, they are exposed to a large number of interviewers, and because of this exposure, they can synthesize the experiences they observe across the entire staff rather than focusing on one person's view. Supervisory staff also are more likely to be familiar with the survey research process and specific survey objectives and therefore are able to sift through comments in a way that results in the identification of common problems or issues that need addressing.

Interviewer debriefing notes can provide a rich narrative on the real-world administration of the survey instrument and provide insight into the data collection process that project staff and survey designers are usually unable to directly observe. Project managers, clients, or survey sponsors can use interviewer debriefings along with other observations gathered from project staff members and respondents and the quantitative survey data collected to identify any problems with the survey effort, questions that may need revision before the next survey administration round, or procedures or processes that need adjustment to meet the goal of fielding the best survey effort possible with the resources available.

Lisa Carley-Baxter

See also Debriefing; Focus Group; Interviewer; Interviewer Effects; Questionnaire; Questionnaire Design; Respondent

Further Readings

Campanelli, P. C., Martin, E. A., & Rothgeb, J. M. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. *The Statistician*, *40*, 253–264.

DeMaio, T. J., Bates, N., Willimack, D., & Ingold, J. (2006). *Supporting document A: Questionnaire testing and evaluation methods for censuses and surveys. Census Bureau Standard: Pretesting questionnaires and related materials for surveys and censuses* (Version 1.2). Retrieved March 29, 2008, from http://www.census.gov/quality/S11-1_v1.2_Test&Eval_Methods.htm

Interviewer Effects

In many surveys, interviewers play an important role in the data collection process. They can be effective in gaining cooperation of the sample persons, helping clarify survey tasks, or motivating the respondent to provide complete and accurate answers. Thus, interviewers can contribute to data quality, but they can also contribute to measurement error. Interviewers can affect respondents' answers through their mere presence as well as their behaviors when administering the survey.

There are several ways in which interviewers seem to influence respondents' answers. First, the presence of an interviewer can stimulate respondents to take social norms into account when answering a survey question. Pressure to conform to social norms can lead to the underreporting of socially undesirable behavior and the overreporting of socially desirable behavior. Second, observable interviewer characteristics, such as age, gender, or race, can affect many stages of the answer process, for example, by changing the salience of the question topic and therefore altering the retrieval process or by influencing the respondents' judgments of which answers would be socially appropriate. Third, the interviewer's verbal and nonverbal behavior can also affect respondents' answers. For example, the interviewer's feedback, facial expressions, or rate of speech can be taken by respondents as reflecting (dis)-approval of their answers or how important the interviewer thinks the question is. Finally, the interviewer can make errors when delivering and recording the answers to a question. These errors are particularly problematic if they are systematic, for example, not reading certain questions exactly as worded, delivering them incorrectly, omitting necessary probes, or neglecting some response categories.

It is important to note that the effects of interviewers on respondents' answers are not equally strong across all types of questions. Social norms apply only to certain behavioral and attitudinal questions. Interviewers' observable characteristics play a role only if they are related to the question content. Early studies on interviewer effects have shown, for example, race-of-interviewer effects for racial items and gender-of-interviewer effects in gender-related attitude questions but no effects with attitude questions related to other subjects. Similarly, the effects of interviewer behavior also vary by question type. They are more likely to occur if respondents are forced to answer questions about unfamiliar topics, questions about topics that are not salient, questions that are difficult to understand, or questions that leave room for differing interpretations to be elicited by the interviewer. Interviewer errors in question delivery are more likely to occur for longer questions or questions asked in series. Filter questions with long follow-up sequences can provide the opportunity for an interviewer to shorten the questionnaire, even when that is not what the researcher wants to happen.

Interviewer effects can have different consequences for survey estimates. Survey researchers differentiate between systematic interviewer effects that bias survey results and variable interviewer effects that increase the variability of a survey statistic while not introducing bias. Results will be biased if most respondents or certain subgroups systematically deviate in the same direction from the "true" score when interviewed by interviewers with specific characteristics or behavior. Race and gender are examples of such characteristics. But interviewer effects can also increase the variance of a survey statistic without introducing any systematic bias into the estimates. If, for example, interviewers have idiosyncratic ways of phrasing a question or conducting probing, all respondents interviewed by the same interviewer will be affected in the same way, but respondents questioned by another interviewer will be exposed to a different set of idiosyncrasies and might alter their answers in a different way. And even if all the individual biases introduced by the interviewers cancel each other out, the interviewers have the effect of increasing the variance of the respondents' answers. This is usually referred to as *interviewer variance*.

In practice, the size of interviewer effects is not only difficult to predict in advance but often even difficult to measure in retrospect. A random assignment of respondents to interviewers (an interpenetrated design) is necessary to estimate the size of such interviewer variance. To assess interviewer bias, knowledge of the respondent's true score or some similar validation measure is necessary. Past research on interviewer effects for attitudinal items has compared answers from groups of respondents randomly assigned to specific interviewers. All that can be learned from those settings is that a pairing of certain types of interviewers and respondents results in answer distortion. However, estimates of the size of this bias and the accompanying answer quality cannot be made without making additional assumptions.

Reducing Interviewer Effects

Several approaches can be taken to reduce interviewer effects.

1. Interviewer effects on variance and bias vary with survey mode. Evidence from past studies has shown larger effects in face-to-face than telephone surveys, and interviewer effects are, by definition, typically absent for self-administered modes. However, in making a decision to eliminate the interviewer, one should keep in mind that interviewers play other important and positive roles in the survey process, for example, in recruiting hard-to-convince respondents. Thus, removing the interviewer may introduce or increase other types of survey errors.

2. If the biasing effect of an interaction among observable interviewer characteristics, question content, and respondent characteristics is well understood, interviewers and respondents could be deliberately matched in ways known to reduce bias. However, even if it were known that certain interviewer–respondent pairings provided more accurate answers, deliberate matching would not be feasible for most surveys because respondent characteristics may not be known in advance or due to legal restrictions prevent the hiring of interviewers based exclusively on observable characteristics. Random assignment of respondents to interviewers is therefore often recommended.

3. Interviewer training can help reduce the variability in interviewer behavior. The data collection process can become more standardized if interviewers learn to do several things more systematically:

- Explain the question-and-answer process to the respondent.
- Motivate the respondent to provide high-quality answers.
- Read questions exactly as worded.
- Probe nondirectively.
- Record answers without interpretation, paraphrasing, or additional inference about the respondent's opinion or behavior.

Interviewers also need to learn to interact with the respondent in a way that minimizes the potential for the respondent to infer a preference for any response. Careful and adequate probing can in some instances lead to higher-quality answers, but in general it appears that interviewers exert more effects on self-reports when the structure of the question–answer interaction is not fully scripted by a protocol. Overall, the risk of interviewer effects can be reduced with well-written questions; for example, questions that are easily understood by the respondent do not carry much cognitive or emotional burden and do not leave room for interpretation.

4. Finally, organizational parameters can be set in such a way that they reduce the likelihood of interviewer effects. Supervising interviewers and monitoring their behavior will help ensure that the questionnaire is implemented as intended. The interviewer reward system can be altered so that interviewers are motivated to focus on achieving not only a high number of completed cases but also high-quality data. A reduction in interviewer workload for a particular survey will further reduce the effects of each individual interviewer on the survey results.

However, it is likely that weaknesses in the interviewer–respondent interchange are inherent to human communication. Thus, reduction of interviewer effects through training or changes in organizational parameters may not be completely successful. Current research is therefore focusing on the possibility of including indicators of the mechanisms that produce interviewer effects in statistical analysis of survey results. This approach requires interviewer information to be part of the dataset.

Frauke Kreuter

See also Conversational Interviewing; Design Effects (*deff*); Interpenetrated Design; Interviewer Monitoring; Interviewer-Related Error; Interviewer Training; Interviewer Variance; Intracluster Homogeneity; Measurement Error; Nondirective Probing; Questionnaire Design

Further Readings

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology*, *12*, 389–437.

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Hyman, H. H., Cobb, W. J., Feldman, J. J., Hart, C. W., & Stember, C. H. (1954). *Interviewing in social research*. Chicago: University of Chicago Press.

Interviewer Monitoring

Interviewer monitoring is a process of observing and evaluating the performance of an individual who is conducting an interview to gather survey data. Interviewer monitoring is typically conducted in an effort to reduce interviewer-related measurement error by allowing the researcher to understand where in the interview mistakes are being made, with whom are they being made, and under what circumstances. Interviewer monitoring is also necessary as a deterrent to interviewer falsification. If the interviewer is made aware that he or she will be monitored and is kept blind as to when the monitoring will occur, the temptation to falsify data can be greatly reduced. This entry contains an overview of the role of interviewer monitoring, followed by a summary of the types of data collected while monitoring, the ways monitoring data can be used to improve the quality of surveys, and finally a summary of the monitoring techniques employed in telephone and face-to-face interviews.

Monitoring Different Aspects of Interviewer Behavior

When interviewers are employed to collect data, they play a central role in the success or failure of a survey. The first responsibility of the interviewer is to persuade the respondent to take part in the survey (for face-to-face interviews, this may be preceded by the need to locate the respondent). Persuading the respondent to participate is not a simple task. It requires a unique set of skills to tailor the conversation to the respondent and win his or her trust while maintaining a professional and courteous presence. More often than not, the interviewer will have only a few seconds before the reluctant respondent will break off from the interview. By monitoring the interaction between the interviewer and respondent, information can be gathered to reduce the frequency with which these break-offs occur.

Once the interviewer gains the trust of the respondent and begins to conduct the interview, a different set of skills is required to ensure the respondent answers every question and does so accurately. If the interviewer is not efficient in administering the questionnaire, break-offs may occur before the interview is completed. There also may be questions that a respondent is not comfortable answering. The interviewer must be able to persuade the respondent to answer these questions without offending him or her or biasing the response. Unless the interviewers are being monitored, those managing the survey have no way of maximizing the frequency and quality of the response the interviewers attain.

Interviewer monitoring also may provide study managers with information that allows them to identify potential problem questions, scripts (verbatim), and nonverbatim delivery. While cognitive testing is typically used to revise survey questions before they are put into the field, there is always a nonzero probability that interviewers still will have difficulty administering a question or that a question still may be misinterpreted by respondents. Interviewer monitoring provides an additional mechanism by which to evaluate the effectiveness of the questionnaire in producing the data desired. If the script or question is verbatim, the wording can be modified based on the information gathered while monitoring. If there is a problem with the way a script or question is being delivered, this information can be used to help the interviewer find a more effective delivery.

Types of Data That Interviewer Monitoring Can Generate

The information gathered while monitoring interviewers can be grouped into four basic categories: (1) operational execution, (2) falsification evidence, (3) interview quality, and (4) questionnaire design. Operational execution includes observations of how accurately the interviewer performs tasks that are unrelated to the actual administration of the questionnaire, but may have an impact on the results of a survey. Examples of these types of tasks include accurately recording the interview outcome or setting up a time to conduct the interview when the respondent is not available on first contact. These types of observations are often based on simple counts and are the source for cooperation, refusal, and completion rates among other commonly reported statistics. Falsification evidence includes any attempt on the part of the interviewer to fabricate or omit data. This could be as minor as purposely skipping a question that should be asked or as major as outright making up data for entire questionnaires. The latter is unlikely to occur when the interview is conducted by telephone in a centralized facility because of the increased likelihood of being caught. It is more likely to occur in face-to-face interviews where the interviewer is either unable to locate the respondent or does not feel safe in the area where the respondent lives.

Interview quality contains observations as to how well the interviewer performs while conducting the interview. These types of observations may include assessments of the interviewer's voice quality (e.g., enunciation, pace, volume), how well the interviewer follows the interview protocol (e.g., coding answers correctly, reading verbatim scripts as written, probing when necessary to clarify answers, maintaining neutrality), and how effectively the interviewer interacts with the respondent (e.g., persuades the respondent to participate, addresses concerns, answers questions). Questionnaire design refers to observations about the flow and delivery of the interview that are a function of the way the script was designed. Poor question wording and incomplete logic skips may be uncovered while monitoring.

How Interviewer Monitoring Can Improve Survey Quality

The data gathered while monitoring interviewers can be used to (a) coach interviewers to improve their performance, (b) inform managers of production issues to be corrected while the study is in the field, (c) develop individual performance assessments that can be used for merit review, (d) make refinements to the script or questionnaire, and (e) assess the amount of interviewer error associated with the given study. Data used for coaching is typically information related to interview quality. These data usually are based on observable behavior that can be objectively assessed. The data should include specific examples of what the interviewer is doing or not doing well and give the interviewer concrete examples of how to improve their performance. Interview quality and operational execution and data falsification are considered core observations that should be monitored throughout the course of a study. Monitoring allows production managers the ability to identify problems when they occur and correct them while the study is under way. These types of observations also make up the core of many performance evaluations. Therefore it is critical that the observations be accurate and reliable.

While some types of errors observed while monitoring interviewers are directly linked to data quality (e.g., failing to complete the interview, miscoding the respondent's answers, and failing to probe and/or clarify ambiguous answers), it is not always clear how other types of errors may impact data quality (e.g., breaking neutrality, failing to read scripts or questions verbatim). Thus the link between interviewer monitoring and the measurement of interviewer-related error in a given survey can often be quite complex. Many measures of interviewer-related error in surveys come from statistical analyses of the survey data themselves rather than the data gathered while monitoring interviewers. These are often reported in conjunction with the production-based observations mentioned earlier.

Interviewer Monitoring Techniques

The ways in which interviewer monitoring takes place vary greatly depending on the type of interview being conducted. For telephone interviews it is common for the person monitoring to listen in unobtrusively from a remote location. This has the advantage of allowing the interview to be monitored in real time, thus reducing the lag between the interview and feedback to the interviewer. It does, however, place limits as to the depth of observations that can be made. In contrast it is very uncommon for a face-to-face interview to be monitored in real time. There are some computer-assisted personal interviewing applications that allow face-to-face interviews to be recorded so they can be reviewed after the interview is complete. This creates a longer duration between the time of the interview and when the interviewer can be given feedback on his or her performance. The advantage of monitoring recorded interviews after the fact lies in the depth at which they can be evaluated. In the end, the nature and extent of interviewer monitoring on a survey is often dictated by the way the survey is being conducted, the type of information needed, and the resources available to devote to monitoring.

Kenneth W. Steve

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Interviewer Debriefing; Interviewer Effects; Interviewer Monitoring Form (IMF); Interviewer Productivity; Interviewer-Related Error; Interviewer Training; Interviewer Variance; Survey Costs

Further Readings

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (1991). *Measurement errors in surveys*. New York: Wiley.

Biemer, P. P., & Lyberg, L. E. (1991). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Steve, K., Burks, A. T., Lavrakas, P. J., Brown, K., & Hoover, B. (2008). Monitoring telephone interviewer performance. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 401–422). New York: Wiley.

Interviewer Monitoring Form (IMF)

An interviewer monitoring form (IMF) is a framework or set of guidelines used to facilitate interviewer monitoring. A carefully developed IMF is central to the development of standardized processes for measuring interviewer-related error and conducting interviewer training and debriefing. The primary purpose for developing an IMF is to minimize interviewer effects and increase interviewer productivity. This entry contains a brief summary of what is known about interviewer monitoring forms, why they are used, and what they should include.

It is widely accepted that interviewer behavior can be a significant source of measurement error in surveys and that effort to observe and reduce this error is a necessary element of conducting survey research whenever interviewers are used to collect data. Interviewer monitoring forms are typically used to monitor the performance and behavior of telephone interviewers. This is due to the fact that telephone interviewers are centrally located and can be observed from a remote location while the interview is taking place. Traditionally it has been up to the individual organization conducting the research to develop its own IMF. Over the past 40 years or more, a wide variety of methods have been developed.

Although many different approaches to monitoring the performance of telephone interviewers have been proposed, there currently exists no standard, widely accepted IMF through which interviewer performance can be assessed and compared across interviewers, studies, or research organizations. This is due in large part to the complexity of the interview process and the wide variety of purposes for conducting telephone interviews. It also, in part, reflects the diversity of "real-world" motivators that influence the nature of the IMF that an individual organization might develop. Still, there exists a growing body of research into understanding what an IMF should contain and how it should be used.

Early interviewer monitoring forms were designed with the sole objective of identifying interviewer errors. This focus stemmed from the desire to quantify the amount of interviewer error a given study might contain. As the industry's understanding of interviewer behavior has become more refined and computer resources have increased, interviewer monitoring forms have become more sophisticated. More and more is being done to better understand the verbal behaviors that contribute to a successful interview, to quantify the behaviors that are observed, and to use these observations to influence future interviewer behavior through feedback and training. Many current IMFs are designed to accomplish multiple objectives. These may include the facilitation of coaching and training, the generation of data for interviewer performance reviews and methodological investigation, and the provision of a framework for interview process auditing.

The content of the IMF depends not only on these objectives but also on the purpose of the interview or the study as a whole. Every study that uses interviewers to gather data should have policies and procedures that the interviewer is expected to follow (e.g., reading certain questions verbatim, coding answers accurately, and maintaining neutrality). These behaviors do not typically produce a great degree of variability because they are baseline expectations. In other words, they are minimum standards that every interviewer should meet as they are central to the effort of reducing interviewer-related measurement error. As such, they commonly serve as the backbone for the auditing process. An IMF may also contain speech-related behaviors that reflect the way the interviewer uses her or his voice or interviewer–respondent interaction behaviors that reflect the way an interviewer deals with reluctant respondents to maintain or promote cooperation.

Examples of speech-related behaviors include enunciation (clarity), modulation (volume fluctuation), and speed (rate of speech). These behaviors are relatively easy to operationally define and observe. Although they may be more powerful predictors of a successful interview, interviewer–respondent interaction behaviors often are more difficult to operationalize. Not all respondents are the same with regard to their willingness to participate. A good IMF will recognize this by containing respondent behaviors that are good indicators of a respondent's willingness, or lack thereof, to participate. A good IMF will also contain behaviors that reflect an interviewer's ability to deal with reluctant respondents. Observing these interviewer and respondent behaviors is critical to maximizing the impact of interviewer training and feedback as they have the highest likelihood of reducing nonresponse and other nonsampling error.

Kenneth W. Steve

See also Interviewer Debriefing; Interviewer Effects; Interviewer Monitoring; Interviewer Productivity; Interviewer-Related Error; Interviewer Training; Interviewer Variance; Nonresponse Error; Nonsampling Error

Further Readings

Cannell, C., & Oskenberg, L. (1988). Observation of behaviors in telephone interviews. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols, & J. Waksberg (Eds.), *Telephone survey methods* (pp. 475–495). New York: Wiley.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, et al. (Eds.), *Survey measurement and process quality* (pp. 287–310). New York: Wiley.

Fowler, F. J. (1991). Reducing interviewer-related error through training, supervision, and other means. In P. P. Beimer, R. M. Groves, L. E. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 259–278). New York: Wiley.

Steve, K., Burks, A. T., Lavrakas, P. J., Brown, K., & Hoover, B. (2008). Monitoring telephone interviewer performance. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 401–422). New York: Wiley.

Interviewer Neutrality

Interviewer neutrality occurs when an interviewer provides no indication of desired responses (remains unbiased) during the interview process. Interviewers are trained to betray no opinion about survey questions to minimize interviewer-related error that occurs when responses are influenced by respondent perception of what the interviewer indicates is an appropriate answer. The process of collecting data using interviewers is designed to obtain valid information (i.e., a respondent's accurate responses), but to be effective the information must be collected in a consistent and neutral manner that minimizes bias. Neutral administration of surveys requires the training of interviewers to not reveal their own opinions or preferences, either verbally or nonverbally, which could induce respondents to provide inaccurate answers in response to perceived interviewer preferences. Specifically, rapport-building behaviors, interviewer feedback, and respondent vulnerability to social desirability and acquiescence need to be considered.

Interviewer neutrality can be accomplished by training interviewers to gather data in a nonjudgmental manner and to use a normal tone of voice throughout the interview process. It is important that interviewers avoid using words or nonverbal cues that imply criticism, surprise, approval, or disapproval. Verbal behavior such as "Yes, I agree" or "I feel the same way," or nonverbal behavior such as smiling, frowning, giving an intense look, or an extended pause may be interpreted by the respondent as approval or disapproval of an answer. Although interviewers are encouraged to establish rapport with respondents to promote respondent motivation, the interviewer must be continually aware of the risk of expressing personal opinions or preferences. When interviewers provide verbal or nonverbal feedback throughout the interview, it is vital to avoid using any feedback techniques that may be interpreted as approval or disapproval. Interviewers should avoid expressing personal opinions on the topics covered in the survey, as well as communicating any personal information that the respondent may use to infer the interviewer's opinions. The validity of the data can be threatened if respondents are aware of interviewer opinions or preferences. Because the goal of interviews is to provide an environment in which respondents feel comfortable reporting accurate answers, it is critical that the interviewers' opinions or preferences do not influence the respondent in any way.

Finally, social desirability (wanting to provide socially acceptable answers) and acquiescence response bias (the tendency to agree with perceived interviewers' opinions) can threaten the validity of the data. Social desirability bias occurs when respondents answer questions to present themselves in a favorable light (providing answers they feel are most socially approved). Acquiescence response bias occurs when respondents agree with statements from the questions that are spoken by the interviewer regardless of content and can lead to responses that merely reflect agreement with what the interviewer is reading rather than the respondent's own opinions. Training for interviewer neutrality should seek to minimize the effects of social desirability and acquiescence. If interviewer neutrality is ensured, interviewer-related error is reduced and thus allows for a more accurate measurement of a respondent's true scores in the variables of interest.

Mindy Anderson-Knott

See also Acquiescence Response Bias; Bias; Interviewer-Related Error; Nonverbal Behavior; Respondent–Interviewer Rapport; Social Desirability; Validity

Further Readings

Bourque, L. B., & Fielder, E. P. (2003). *How to conduct telephone surveys* (2nd ed.). Thousand Oaks, CA: Sage.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Eds.). (2004). *Survey methodology*. New York: Wiley.

Oishi, S. (2003). *How to conduct in-person interviews for surveys* (2nd ed.). Thousand Oaks, CA: Sage.

Interviewer Productivity

Interviewer productivity refers to the ways of measuring what is achieved by telephone and in-person survey interviewers when they work to (a) gain cooperation from sampled respondents and (b) complete interviews with these respondents. Measuring interviewer productivity is a major concern for those conducting and managing surveys, for several reasons. Knowledge of productivity is essential to survey budgeting and developing realistic estimates of survey costs. Managing a survey requires an understanding about how many completed interviews, refusals, noncontacts, ineligibles, and callbacks can be expected for a given survey. Productivity information is often used to reward interviewers that are performing well or to retrain those who are not being productive (enough). Interviewer productivity information is also a necessary aspect of planning and scheduling the number of interviewers needed for fielding a survey and for monitoring survey progress. It is also important to communicate productivity expectations to interviewers in advance of the start of data collection so they know how to perform adequately.

Literature

Interviewer productivity is discussed in the literature on survey methods. Don Dillman has noted the importance of interviewer productivity by indicating that the length of time taken to complete interviews needs to be taken into account in scheduling interviewers. Paul Lavrakas has proposed that interviewer productivity be measured by the speed with which samples are worked and also by the ratio of the number of completed interviews to the number of refusals and partially completed interviews. He suggests using these kinds of productivity measures to set interviewer pay rates and manage surveys. Interviewers who are more productive also have a greater influence on the amount of survey data collected than do less productive interviewers. This is because more productive interviewers end up completing more interviews than less productive interviewers. Robert Groves describes how estimates of the standard errors in a survey are directly related to the average number of interviews completed by each interviewer. In contrast, interviewers who are less productive may well have greater influence on other aspects of survey data quality such as nonresponse and possible nonresponse error.

Influences and Uses

There are currently no agreed upon standards of interviewer productivity because productivity can be influenced by so many characteristics of an interviewer-administered survey, including the survey mode (telephone, face-to-face), the survey population, the length of the questionnaire, the experience of the interviewer, and the particular phase of survey contact (initial contacts vs. callback attempts). The most frequently used measure of interviewer productivity is the number of completed interviews obtained by an interviewer during some set period of time, such as "per hour." This is important to measure, as most surveys have a goal to achieve a specific number of completed interviews within a pre-specified field period length.

Interviewer productivity can be influenced by incentives added to their compensation. At least one study has found that the number of interviews an interviewer will complete during a shift can be affected by the offer of a monetary reward pegged to productivity. However, rewarding interviewers for productivity has potential drawbacks (e.g., interviewers may engage in cheating to earn monetary rewards), so this practice is generally frowned upon. Thus, whenever an incentive system is considered for rewarding interviewers for their productivity, it is critical that it be combined with a reliable and rigorous system of monitoring the interviewers.

Computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI) are the two main survey modes that use interviewers and that are concerned with interviewer productivity. However, CATI surveying is the predominant survey mode for most survey organizations, and the majority of productivity measures that have been developed are those obtained with CATI systems, which facilitate the collection of many measures of interviewer productivity, including (a) number of call attempts per hour, (b) number of completed interviews per hour, (c) number of refusals per hour, (d) number of ineligibles per hour, (e) total number of minutes worked, (f) average length of completed interviews, (g) quality of interviews completed, (h) monitoring scores, and (i) attendance and tardiness. Some of these measures clearly apply to face-to-face interviewing with CAPI.

Information on interviewer productivity is generally used by survey managers to forecast progress toward survey deadlines and goals and also for communicating with interviewing staff about their performance on a survey. Interviewer productivity is often used to estimate the number of interviewers needed to meet survey goals or the number of hours or days required to reach a survey deadline. For example, if interviewer productivity for a particular survey is measured as an interviewer taking a half hour to complete one interview, and the survey goal is to have 400 completed interviews, then multiplying 400 by one half hour equals 200 interviewer hours that are required to reach the survey goals. However, since hours per complete interview tends to increase while a survey is in the field—because of an increasing proportion of unproductive calling (no answers, answering machines, refusals)—it is usually more accurate to recalibrate interviewer productivity periodically throughout the survey and re-estimate the number of interviewer hours needed to reach survey goals. This calculation can be made easier, for example, by logging daily productivity information on a spreadsheet, which can then be used to forecast the number of interviewer hours and the number of days required to meet survey goals.

Performance Evaluations

Interviewers want to know how their performance is evaluated, and productivity information is a useful way to provide this kind of performance evaluation. Survey organizations often post interviewer productivity information for every interviewer on a survey, such as the number of calls per hour, hours per completed interview, refusals per hour, and ineligibles per hour that interviewers achieve over the course of a survey project on a daily basis. Because the productivity of every survey is different, this information provides interviewers, supervisors, and managers with a way to assess their progress on a particular survey. Some organizations calculate control limits on measures of interviewer productivity and then identify which interviewers are outside the acceptable range of performance. Interviewers who are outside the range of the control limits on the high side may be promoted (e.g., to become lead interviewers, monitors, or supervisors), whereas others may be investigated to determine if there is a possibility of falsification. Those outside the range of control limits on the low side may be sent for additional training to improve their productivity. However, productivity measures are generally correlated, so they must be used carefully by survey management. For example, interviewers who achieve a high rate of completed interviews will generally see their calls per hour decrease because their time is occupied with interviewing respondents, which takes more time than making call attempts.

Future Directions

Most survey organizations are concerned about interviewer productivity, including declining response rates, lack of trained interviewers, and the problems of managing a part-time work force of interviewers. A survey of telephone survey organizations conducted in 2007 found that almost all survey call centers (84%) regularly collect productivity information on interviewers. However, only 54% reported that they actively use this information as a way to analyze and to make decisions about interviewers. Standards of interviewer productivity differ among survey research organizations and are dependent on the specific goals of each organization and of each survey. But no survey organization can survive if it does not pay attention to and regularly measure interviewer productivity. In the future, interviewer productivity measures will become even more important, as survey costs increase and organizations look for ways to increase ever-declining survey response rates.

John Tarnai and Danna L. Moore

See also Completed Interview; Completion Rate; Computer-Assisted Telephone Interviewing (CATI); Face-to-Face Interviewing; Falsification; Field Period; Field Work; Interviewer Effects; Interviewer Monitoring; Mode of Data Collection; Response Rates; Supervisor; Survey Costs

Further Readings

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Frey, J. H. (1983). *Survey research by telephone*. Beverly Hills, CA: Sage.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods* (2nd ed.). Newbury Park, CA: Sage.

Steve, K., Burks, A. T., Lavrakas, P. J., Brown, K., & Hoover, B. (2008). Monitoring telephone interviewer performance. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 401–422). New York: Wiley.

Tarnai, J., & Moore, D. L. (2008). Measuring and improving telephone interviewer performance and productivity. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 359–384). New York: Wiley.

INTERVIEWER-RELATED ERROR

Interviewer-related error is a form of measurement error and includes both the bias and the variance that interviewers can contribute to the data that are gathered in face-to-face and telephone surveys. In interviewer-administered surveys, although interviewers can contribute much to the accuracy of the data that are gathered, they also can contribute much of the nonsampling error that finds its way into those data.

The methodological literature includes startling examples of measurement error due to interviewer mistakes. In 1983, an interviewer's incorrect recording of one wealthy respondent's income resulted in the erroneous report that the richest half percent of the U.S. population held 35% of the national wealth. This finding, widely publicized, was interpreted to show that Reaganomics favored the wealthy. When the error was detected and corrected, the actual estimate was 27%—only a slight increase from the 1963 figure. Most survey designs do not feature weighting schemes that permit one interviewer's random error to have such a profound effect. Usually, random interviewer errors "cancel each other out," not threatening data validity.

It is systematic, rather than random, interviewer-related error (i.e., bias) that typically affects survey data. Systematic, or correlated, error occurs when interviewers make similar "mistakes" across many interviews. Such errors may actually reduce item variance, but they play havoc with the accuracy of resulting estimates. This entry focuses on the sources of, and treatments for, systematic interviewer error and discusses efforts to prevent, measure, manage, and correct for this type of bias.

Preventing Interviewer-Related Error

Prevention focuses on three basic strategies: (1) reducing or eliminating human intervention between respondent and data capture, (2) engineering error-proof questionnaires and associated data collection tools, and (3) standardizing interviewer behaviors to minimize error.

In their review of interviewer-related error, Robert M. Groves and his colleagues note that the very presence of an interviewer has been shown to bias responses. Employing computerized, scanned, or voice-response self-administration avoids both the costs and errors associated with employing human interviewers. Sloppy data resulting from respondent-related error, the bugaboo of self-administration, can be attacked through programming that mandates response and requires clarification of contradictory information.

This approach, however, has its own drawbacks. Notable among them are higher front-end costs and lead time, limitations on complexity, and lower response rates. Cybernetic approaches are expensive for all but the simplest of questionnaires; they require intensive programming, pretesting, and debugging to meet the requirement that they be far more error-proof than is required when a trained interviewer is involved. Even for simple questionnaires, the value added by the presence of an interviewer to motivate engagement and probe for focused and detailed answers compensates for and usually exceeds the error contributed. Complex enumerations and life history matrices are approached with trepidation in the absence of trained interviewers. Finally, it is far easier for respondents to opt out of higher-burden self-administered surveys than to avoid or disappoint a pleasant, yet determined and persistent interviewer.

Because both interviewer-administered and self-administered data collection have strengths and weaknesses, in surveys where the elimination of interviewers entirely is not prudent or possible, questions known to be affected by interviewer characteristics or limits to their capabilities can be switched to self-administration. Barbara Mensch and Denise Kandel found in their reanalysis of data from a longitudinal study that young respondents having the same interviewer over multiple data collection waves significantly underreported drug use. Their conclusion was that the development of "over-rapport" with the interviewer heightened self-censorship. Because that very rapport was responsible for the panel's extraordinary retention (ongoing response) rate, the solution was to maintain continuity of interviewer assignments but move the sensitive questions to self-administration.

A second major strategy for preventing interviewer-related error is engineering error-proof questionnaires and associated data collection tools. Sometimes interviewers slip up; it may be due to principal investigators laying out "banana peels" in the form of confusing questions, excessive burden, "overly academic" language, complex skip instructions, and crammed layouts. The burgeoning literature on designing Web surveys for clarity and ease of self-administration provides guidance for improving all surveys—these questionnaire design principles, adapted to interviewer-administered questionnaires, also increase an interviewer's ability to provide error-free data. Pretesting, focus groups, and cognitive interviews lead to questionnaires that have fewer opportunities for error.

The third prevention strategy is standardizing interviewer behavior to censor actions associated with measurement error. Two competing theories of how best to do this suggest different approaches to preventing interviewer error. The dominant paradigm seeks to severely constrain interviewer behavior as a means to *standardize the presentation of the stimulus* by the interviewer across all interviews. An alternative conversational approach to interviewing seeks to *standardize the understanding of the stimulus* by the respondent across all interviews—that is, to communicate what the investigator really wants to know, even when the question must be adapted to the particular respondent's cognitive abilities, language skills, and knowledge.

In the standardizing-the-stimulus approach, questions must be asked exactly as written in the order provided; only neutral probes can be used and only scripted answers can be offered for respondent questions; and interviewers can offer a severely restricted set of verbalizations, none of which is leading or even affirming. Ad-libbing is anathema, sometimes even on scripted introductory materials. These "unfriendly" and sometimes counterintuitive standards can contribute mightily to break-offs (partial completions), as many respondents grow weary of hearing the same answer categories repeated or are forced to listen to and answer a question they think they have already answered. But the cost is seen as a worthwhile trade-off for the benefits of keeping tightly lidded the "Pandora's box" of interviewer discretion.

Practical experience and the literature note sizable variability in interviewer compliance with these principles. For example, Peneff found widespread violation of norms in his observational study of French interviewers. For this and because of the tyranny of conversational norms, interviewers may try to smooth awkward situations in ways that cause respondents to censor their remarks.

Field studies use interviewer observations and telephone survey calling centers use random monitoring to evaluate and code interviewer behavior, but the vast majority of survey interviewing is unobserved in both telephone and face-to-face surveys and thus prone to error introduced by noncompliance with strict interviewing rules. Past research has indicated that interviewers have been found to change the wording from 20% to 95% of the time and that interviewer experience (tenure) is uncorrelated with this form of error. Training is seen as one of few available responses, but training itself has not been found to be highly correlated with subsequent reductions in error. In many cases, wording changes are minor and do not affect the answer. Failure to probe and inadequate probing have been found to be major contributors to nonnegligent interviewer error.

In the standardizing-the-meaning paradigm (also called conversational interviewing or flexible interviewing), interviewers are given discretion to vary the script if they judge it will increase response accuracy. In a telephone survey experiment conducted in the late 1990s, Fred Conrad and Michael Schober found that this approach increased valid answers on ambiguous factual questions. Standardizing-the-meaning approaches result in longer interviews and higher costs due to the extra time in training interviewers about the goals of specific questions and the subsequent extra time during negotiating question meaning with respondents. This paradigm has been in use for decades but has been mostly limited to elite interviews and in situations where a structured questionnaire is viewed as inappropriate or ineffective.

Flexible interviewing opens the door to untrammeled interviewer discretion, and once opened, it may be difficult for interviewers to identify those situations that require the standardizing-the-stimulus approach and those that do not. On the other hand, the counterintuitive restrictions on interviewer behavior in the standard model foster violation of rules, which could bleed over into other aspects of interviewing. One solution to this dilemma is to designate certain key questions whose wording is ambiguous in meaning to respondents as standardizing-the-meaning and clearly restrict interviewer discretion to these questions only. In effect, that is what likely occurs in the field in many cases, as interviewers left to their own devices try to negotiate better answers from confused or recalcitrant respondents.

Measuring Interviewer-Related Error

Although interviewer-related error cannot always be prevented, it usually can be measured. Measurement can then lead to decisions about managing away future error, correcting the tainted data, or ignoring it. The gold standard in measuring interviewer variance is the intraclass correlation, rho. Rho involves calculating an analysis of variance with the interviewer as the treatment. Rho measures how much of the variance is contributed by interviewers; even a small rho can sizably inflate the variance of a variable. Unfortunately, the practicalities of survey work do not play to the strengths of rho. Interviewers often cannot be assigned randomly to respondents, especially when face-to-face interviewing is the mode of data collection; in current survey operations, multiple interviewers handle a single case and specialists may be assigned all the cases of a particular type. Furthermore, one telephone interviewer may handle 50 cases while another completes 2 cases. These conditions either violate rho's assumptions or limit its practical applicability.

A second approach to measurement is to create an ongoing "error profile" from the collected data. Questions central to the anticipated analysis are tracked in terms of missing data, "don't know" responses, data reduction staff assessment of the quality of responses to open questions, and any other more easily quantified measure related to error. Some questions will show high error rates across all interviewers. These are the "banana peels" in the questionnaire, best addressed by redesign and retraining across-the-board. Others will show high variance in error rate by interviewer. These are the interviewer-specific correlated errors that can be addressed through targeted retraining or more draconian solutions.

The most common approach to measurement of interviewer-related error involves observing field interviewing and monitoring telephone interviews. Interviewer observation can cover only a minimum of actual interview time (even in telephone centers) and is expensive, labor intensive, and difficult to analyze in real time. Quantitative coding of interviewer behavior is used in telephone survey labs to offer corrective feedback, but, in part because of the disconnect between day-to-day managers and principal investigators, there is little evidence that these protocols protect the integrity of key research hypotheses.

Managing Interviewer-Related Error

Training, supervision and ongoing feedback, interviewer–respondent matching, and validation are four ways to manage interviewer-related error. Jack Fowler and Tom Mangione have provided most of what little empirical evidence exists in the area of training and supervision. Training and supervision are shown to be effective in reducing error; however, more experienced interviewers are sometimes more likely to make errors than those with less experience—perhaps because they feel they have earned the right to cut corners and use their discretion.

A very small number of interviewer characteristics are proven causes of bias. Herb Weisberg provides a useful review of this literature, concluding that surveys directly related to interviewer characteristics are ones to worry about, such as anti-Semitism questions asked by Jewish interviewers. Gender effects have been noted but are difficult to generalize. Unambiguous race effects on questions that deal with racial matters lead to the rule of thumb to race-match in such interviews. Managing interviewer effects usually amounts to matching interviewer and respondent characteristics as closely as possible, "just to be safe." Weisberg's counterpoint is that it is not known if matching race results in overreports of race-biased attitudes and interpretations. Much of the work done in this area is decades old.

Interviewing invalidation, that is, fabricating answers to part or all of the questions interviewers ask, is a dark cloud on the horizon of survey research. A determined interviewer who falsifies data makes detection almost impossible. Management efforts to uncover validation problems include reinterviews with respondents (often difficult to complete or sadly ambiguous), analyzing the data by interviewer to search for suspicious patterns (hobbled by small sample sizes), and careful follow-up on questionable interviewer behavior.

Correcting Interviewer-Related Error

Most interviewer-related error cannot be detected and thus cannot be corrected. The first line of defense is increasingly foolproof questionnaires, better training, and closer supervision. The second is careful analysis of missing and suspect data and, where possible, imputation of missing data. The greatest effect in reducing and correcting error would stem from increased research into the sources and solutions for interviewer-related error, a research area generally neglected in contemporary survey methods research.

Woody Carter

See also Bias; Cognitive Interviewing; Conversational Interviewing; Face-to-Face Interviewing; Falsification; Focus Group; Interviewer Characteristics; Interviewer Effects; Interviewer Monitoring; Interviewer Neutrality; Interviewer Training; Measurement Error; Nondirective Probing; Nonsampling Error; Partial Completion; Pilot Test; Probing; Reinterview; Respondent-Related Error; ρ (Rho); Self-Administered Questionnaire; Standardized Survey Interviewing; Telephone Surveys; Validation; Variance; Verification

Further Readings

Conrad, F., & Schober, M. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, *64*, 1–28.

Ericksen, E. P. (1988). The polls—a review: Estimating the concentration of wealth in America. *Public Opinion Quarterly*, *52*, 243.

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Groves, R. M. (1989). *Survey error and survey costs*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology* (pp. 269–298). Hoboken, NJ: Wiley-Interscience.

Mensch, B. S., & Kandel, D. B. (1988). Underreporting of substance use in a national longitudinal youth cohort: Individual and interviewer effects. *Public Opinion Quarterly*, *52*, 100–124.

Peneff, J. (1988). The observers observed: French survey researchers at work. *Journal of Social Problems*, *35*, 520–535.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

Interviewer Training

Interviewer training refers to the instruction that survey research interviewers receive at various stages of their employment, and in various ways, to make it more likely that they will perform their jobs effectively. It is absolutely essential for achieving high-quality survey samples, interviews, and resulting data. Organizations that hire people to conduct standardized survey interviews understand that one of the most critical success factors is how well interviewers are trained and managed. The purpose of training interviewers is to teach the principles, skills, and basic procedures needed to conduct telephone or face-to-face interviewing in a manner that achieves high-quality, reliable, and valid information for research.

Training Elements and Expectations

Initial training for new interviewers provides the fundamentals, such as the nature of interview work, workplace ethics (e.g., attendance, professionalism), using computer-assisted interviewing (CATI or CAPI) systems, and emphasizing standardized interviewing techniques. Procedures for training interviewers should ideally focus on all the skills for basic interviewing needed across most surveys in general.

For maximum effectiveness, interviewer training must convey performance expectations for interview work and teach interviewers how to conduct a standardized interview. Interviewer training describes acceptable methods for questioning respondents and how to collect interview information with a high degree of accuracy and reliability. Because interviewers are the intermediaries of questionnaires (and questions) to respondents, they can also be an important source of error in the survey process. The crux of interviewer training then is to reduce interviewer-mediated error. This is accomplished by educating interviewers and demonstrating proper interviewing techniques, such as how to read questions as worded, neutrally probe respondents, relate to respondents so as not to introduce bias to survey answers, and record answers as accurately as possible. Overall, trainings should reinforce interviewers' interactions with respondents as interpersonally neutral while asking interview questions. Interviewer training should not only introduce and explain basic interviewing concepts but also provide time for deliberative practice of skills. Trainees gain knowledge, skill, and confidence through observation and participation in activities that mimic real interviewing.

Basic and Project-Specific Training

In a 2007 study of telephone survey research organizations, John Tarnai and Danna Moore found that almost all such organizations use trainings to increase interviewing skill and knowledge and that new interviewers are required to participate in training before they can begin calling. Most organizations have developed trainings that include written procedures for standardizing interviewing and on average spend about 6 hours per interviewer to cover basic interviewing training and practice. Table 1 displays the main topics and activities regularly covered in basic introductory trainings by telephone survey organizations (of note, these same topics and activities essentially apply to face-to-face interviewing).

Many organizations also regularly hold project-specific trainings, which detail requirements and circumstances on individual surveys. Trainings may also be expanded to cover complex survey situations; for example, questionnaires that have complex conditional skip patterns or that screen for eligible survey respondents often require additional training for interviewers. Other examples of the need for project-specific training are when the survey topic is sensitive or includes respondents classified as members of a vulnerable or protected population. Both of these situations will generally require talking about human subject research, confidentiality, the rights of respondents, and special issues raised by respondents.

Table 1 Main interviewer training topics

- Description of standardized interviewing and the rules to adhere to
- Explanation and examples of acceptable probing and feedback
- Demonstration of how to use CATI software systems
- Explanation of the various types of questions used in data collection
- Explanation of respondent reluctance and cooperation
- Proper interaction with respondents
- Addressing respondent concerns
- Explanation of survey introductions
- Practice and role play by interviewers
- Explanation of refusal behavior and interactions
- Explanation and use of case disposition codes
- Practice and role playing mock interviews
- Data entry errors and ways to make corrections
- Explanation of research ethics
- Expectations of interview performance and how performance is measured
- Explanation and practice of dialing
- Human subjects research and confidentiality
- Special emergency situations and what to do

Additional Types of Training

There are other types of trainings for interviewers such as those that specifically address performance improvement or advanced interviewing tasks, such as refusal conversions or addressing and diagnosing survey problems. These trainings focus on improving individual interviewing skill for increasing survey participation and countering nonresponse at both the survey introduction and the item level. These specialized trainings are intended to increase interviewers' awareness of their own behavior during different parts of their contact with respondents, to help them be more able to recognize respondent's concerns, and to practice delivering rapid responses to concerns raised by respondents. Research on interviewer training has shown that advanced trainings such as refusal avoidance, tailoring introductions, and recognizing and addressing respondent concerns can positively change interviewer behavior and lead to greater respondent cooperation. The distinguishing characteristics of high-performing experienced interviewers are that they (a) engage respondents with more confidence, (b) have a larger combination of behaviors, (c) prolong interactions, and (d) actively communicate to persuade respondents.

Interviewers are often required to conduct practice interviews by pairing-off to role play with one another while supervisors monitor their performance, offer a critique, and provide feedback. In real life, difficulties will be encountered during interviews, and supervisors can prepare interviewers by acting the part of a recalcitrant respondent during trainings. Supervisors may role play the part of a particularly recalcitrant respondent and make suggestions to interviewers about "what to say" and "how to say it" in the most persuasive manner, thus affording interviewers with the tools needed to effectively address the noncompliant respondents they are likely to encounter.

Trainings are useful for observing how interviewers interpret survey and training materials, how they translate knowledge into practice, and for confirming interviewer understanding of what constitutes a standardized interview. Some larger survey organizations have moved toward using computer-based training, interviewer certification, or other testing systems to assess whether new interviewers have sufficient knowledge and training to move to production interviewing.

Interviewer training is significantly related to improved interviewer performance, which most survey organizations regularly measure. Assessment of interviewer performance through monitoring on a regular basis and the use of interviewer-based metrics across surveys provides documentation of whether trainings have a measurable impact collectively on interviewer skills and the accuracy of data collected. Past research has shown that individual interviewer effects (survey errors) in a survey are greater when the number of interviewers working on a survey is small or when any one interviewer's workload (number of completed interviews) is significantly greater than that of other interviewers. It is important to know whether interviewer effects are occurring in a survey and when to adjust for this. The main way to correct interviewer behavior is to provide specialized training for addressing problems or increasing certain skills. Interviewer training and productivity management go hand in hand to influence and direct interviewer skill development and behaviors.

Communicating productivity expectations to telephone interviewers during training provides interviewers with benchmark information on job tasks and provides measures against which they can compare their own level of performance and so they know how to perform adequately or what they need to do to improve. Supervisor interactions with interviewers should reinforce expectations stated during trainings. Productivity plays an important part in interviewer training, as increasing productivity through more completions, fewer refusals, fewer missing items, and more call attempts are generally the goals of interviewer training.

Training interviewers, regularly monitoring their performance, and providing direct and timely feedback to interviewers about their performance is the most consistent way to establish and ensure that interviewing as a process is done to a set standard and according to best practices. Training of interviewers is essential for producing high-quality survey data. Interviewer training introduces and emphasizes quality aspects of interviewing work. For instance, achieving acceptable response rates is important for the reliability and validity of a study and is directly linked to interviewers being able to get respondents to cooperate with the interview. To gain high rates of respondent participation, interviewers need to be taught basic persuasion techniques. Confidentiality is another area of concern to many respondents, and interviewer trainings should emphasize assuring respondents that their identifying information is kept private and that their individual survey responses will not be reported. Effective trainings incorporate a thorough description of what it means to keep data confidential, how data can be compromised, and what its importance is to respondents. At issue is interviewers' ability to alleviate respondent fears that confidentiality cannot or will not be ensured.

Refusal conversion training is another type of interviewer training that is important to survey organizations. Converting respondents who have initially refused interviews has become an increasingly important strategy for achieving acceptable response rates and for bringing in respondents to a study that are inherently different from the early, more easily obtained respondents on measures important to the study results. More formally, this is referred to as trying to reduce nonresponse and possible nonresponse error. Training interviewers to effectively convert refusals requires increasing their knowledge of the reasons respondents refuse and providing interviewers with specific statements and appeals that can be used to counter respondents' refusal language.

Training Standards

Interviewer training is an important way to convey to interviewers what is expected of them and how to adequately perform the job of interviewing in a standardized way. Survey training activities are very important to survey quality and are considered a best practice for survey organizations. The International Standards Organization has established a set of standards for market, opinion, and social research that includes a requirement for at least 6 hours of training for new telephone interviewers. The goal of this requirement is to ensure that all interviewers receive a minimum amount of training in standardized survey interviewing, thereby ensuring better quality survey results.

Danna L. Moore and John Tarnai

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Confidentiality; Interviewer Monitoring; Interviewer Productivity; Interviewer-Related Error; Nonresponse; Nonresponse Error; Probing; Refusal Avoidance Training (RAT); Refusal Conversion; Role Playing; Survey Ethics; Tailoring

Further Readings

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475–495.

Groves, R. M., & Couper, M. (1998). *Nonresponse in household surveys*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Groves, R. M., & McGonagle, K. A. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17*(2), 249–265.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Tarnai, J., & Moore, D. L. (2008). Measuring and improving telephone interviewer performance and productivity. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 359–384). New York: Wiley.

Interviewer Variance

Interviewer variance describes the part of the overall variability in a survey statistic that is associated with the interviewer. Clusters of respondents interviewed by the same person tend to have more similar responses than do clusters of respondents interviewed by different interviewers. This cluster effect can appear, for example, if an interviewer uses inappropriate or inconsistent probing techniques, has idiosyncratic interpretations of questions and rewords them accordingly, or differs in the way he or she reads answer categories. In addition, interviewer-specific interactions between the interviewer and respondent can lead to an intra-interviewer covariance term that contributes to the variance of the estimate.

The effect of interviewers on responses can increase the variability of survey estimates in a way parallel to the effect of clustered samples. The standard errors of such survey estimates are inflated compared to those computed for a simple random sample. Thus, ignoring the clustering of respondents within interviewers can yield misleading results in significance tests or in the coverage rates of confidence intervals. Most statistical packages use linearization or replication methods to correct the variance estimation for different kinds of sampling designs. To account for an interviewer clustering effect, those procedures require either an interviewer identification variable or appropriate replicate weights created by the data collector as part of the data set.

The overall variance of the respondent mean is inflated by interviewer variance according to the function $deff = 1 + \rho(w - 1)$, where w is the average number of interviews conducted by individual interviewers, and ρ is the intraclass correlation coefficient among responses produced by a common interviewer. If all respondents interviewed by the same interviewer answered in exactly the same way, ρ would be equal to 1. The size of ρ reported by various researchers has shown substantial variation among surveys and survey variables. The average value for ρ in many (mostly telephone) studies is 0.01, but values of about 0.05 are not uncommon, while for some surveys and items a ρ as high as 0.2 has been observed. These seemingly small values can have a large impact. If the average workload for an interviewer in a survey is 100, a ρ of

0.01 can produce a design effect of 1.99. Both a high workload and a high value of ρ contribute to a problematic design effect. A value of $deff = 2$ would be equivalent to cutting the sample size in half.

Telephone surveys often have high interviewer workloads. Thus, even with low values for ρ, interviewer variance may be a problem and should be accounted for in the data analysis. In face-to-face surveys, not only interviewer variance but a second source of variance contributes to the size of the design effect. Interviewers often work in small geographical clusters to reduce the cost of data collection. The attributes of respondents interviewed by the same interviewer can therefore be correlated simply because people who live in close proximity are likely to be similar to each other in some way. To determine whether it is the interviewer or the geographical proximity that is responsible for the cluster effect, an interpenetrated sample design is required, one in which a random subsample of the full sample is assigned to each interviewer. In practice, there often are considerable limitations to implementing interpenetrated assignments. They are usually too expensive for area probability samples. And randomizing telephone cases among interviewers can usually be done only within the shifts that interviewers work.

Instead, interviewer variance can be reduced if fewer cases are assigned to an individual interviewer. (Although if too few cases are assigned on average, other sources of interviewer error may be inflated; e.g., interviewers may not become as proficient in administering the question.) Other strategies include measures to enhance the standardization of the interview procedure through training and monitoring of the interviewer. Improved questionnaire design can also help reduce interviewer variance. To correct standard errors for clustering effects due to interviewers, data sets must provide interviewer identification variables, a practice that is still rather uncommon.

Frauke Kreuter

See also Clustering; Design Effects (*deff*); Interpenetrated Design; Interviewing; Replicate Methods for Variance Estimation; ρ (Rho); Standardized Survey Interviewing

Further Readings

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, *11*, 55–77.

INTERVIEWING

Survey interviewing is typically a formal, standardized conversation between a person asking questions (the interviewer) and a person giving answers to those questions (the respondent). The respondents are selected because they belong to a population of interest. The population can be very broad (e.g., residents of a city or state, registered voters) or very narrow (e.g., people who have been diagnosed with a particular disease; females who smoke cigarettes, have less than a high school education, and watch the local news on a specified television station). In addition to asking questions, the interviewers may also play other roles, such as gaining initial cooperation from the respondents or showing respondents how to answer self-administered questionnaires on paper or by computer. While some survey data may be collected by self-administration (e.g., mail or Internet-based surveys), many surveys, particularly long, complicated ones, require the use of an interviewer. Thus interviewing is an important aspect of survey research. This entry provides an overview of factors relevant to interviewing, many of which are discussed in greater detail in other entries in this volume.

A Short History of Interviewing

Survey interviewing, or its equivalent, has been conducted for thousands of years, from ancient times when rulers sent out census takers to gather information about the people they found (in households or elsewhere), including the gender and age of each person. Businesses have long queried their customers and clients about products and services. In the 1930s and 1940s, U.S. government agencies began conducting many more surveys than before. As before, the interviews were conducted using paper questionnaires. This was called PAPI, for paper-and-pencil interviewing. At that time almost all interviewing was conducted face-to-face.

In the United States, telephone interviewing became popular in the 1960s, because by that time most households in the country had telephones. (This was not the case in Europe and other developed countries, where private telephones were very expensive and could take months to get installed.) However, face-to-face interviewing was still used for long, complicated surveys and those that required visual aids or physical tasks such as card sorting.

While large main-frame computers had been used for decades for survey data processing, it was not until the widespread availability of minicomputers—later called personal computers (PCs)—in the late 1970s and early 1980s that interviewing became computerized. Thus computer-assisted telephone interviewing (CATI) was born. Later, when laptop computers were created and became affordable, computer-assisted personal interviewing (CAPI) became common. (At this point, European surveyors took the lead with this technology, having skipped over CATI because of lower residential telephone coverage.)

As more respondents became familiar with them, computers were used for computer-assisted self-interviewing (CASI). Respondents often need assistance with how to answer computerized questionnaires, so typically an interviewer plays the role of instructor. With the introduction of a sound component, audio computer-assisted self-interviewing (ACASI) became possible. It is primarily used to (a) ask sensitive questions that a respondent might not want to answer to an interviewer, (b) ask questions in languages other than the one or two used on the written questionnaire, or (c) offer an oral version of a questionnaire to a respondent who cannot read well.

Interviewer Effects

Interviewing can cause *interviewer effects*, a subset of a larger problem called *measurement error*. The primary causes of interviewer effects are improper administration of the questionnaire and the effects of interviewer characteristics themselves. Interviewers can negatively affect the administration of a questionnaire in many ways, such as by misreading questions and deviating from a standardized script. Improper questionnaire administration can be held to a minimum through the use of professional, well-trained interviewers. The way to minimize interviewer-related error is to standardize the interviewing as much as possible.

Interviewer Characteristics

Interviewer characteristics are those things about a particular interviewer that may, in some circumstances, affect how a respondent will answer survey questions. They include gender, age, race, ethnicity, and in some countries, perceived or actual class, caste, or clan. Interviewer characteristics can affect respondent–interviewer rapport and respondent answers, positively or negatively. In some cases, if there is a major concern that an interviewer characteristic may affect the respondent, interviewers may be matched with respondents on that characteristic. For example, only female interviewers would interview female respondents or only older male interviewers would interview older male respondents.

Interviewer Training

The quality of survey interviewing is dependent on good interviewer training. Surveys conducted by professional survey organizations use professional interviewers who have been trained extensively. At the other extreme, some surveys are conducted by volunteers who have been given little or no instruction. Regardless of whether or not interviewers are paid, it is essential that they are trained to do the job expected of them. While a large part of interviewer training relates to the administration of the questionnaire (e.g., asking the questions as worded; reading or not reading answer categories, as directed; knowing when and how to probe), there are many other important aspects to this training. They include (a) identifying the appropriate sample unit (e.g., a household, a telephone number, an individual), (b) obtaining and maintaining respondent cooperation, and (c) following prescribed rules regarding contacts (number of attempts, days and times of attempts, etc.). In addition to the information in the questionnaire's introduction, interviewers should be given backup information, including "fallback statements," about the study so that they can answer potential respondent questions. Such questions may include which firm is conducting the survey, who is sponsoring or funding the survey, why the household or respondent has been selected, approximately how long the interview will take, whether or not the information will be confidential, and who to contact if they want further information.

Regardless of the length of the questionnaire and the experience of the interviewers, the person conducting the survey or supervising the data collection should prepare a written interviewer training manual. The manual should be used for initial training and also be available for review throughout the data collection period. In fact, if certain instructions, probes, lists, and so forth, are important and may need to be referred to during an interview, they should be either printed out separately or available on the computer screen, so that an interviewer can refer to them instantly.

Most important is to make clear to the interviewers what the "conventions" are on the questionnaire, for the interview. In other words, (a) when do they read a question and stop to hear the respondent answer, and (b) when do they read a question and continue reading predetermined answer categories? This is a very important part of standardization for any survey administration, for it is known that respondents will answer differently if they are given answer choices or categories to select from (closed-ended question) or not (open-ended question). It is important that these conventions be consistent throughout the questionnaire. For example, a question mark after a statement might indicate to stop reading the question and wait for the respondent to answer, while a series of three dots might indicate to continue reading answer categories:

(a) *Do you agree or disagree?*

(b) *Would you say . . . strongly agree, somewhat agree, somewhat disagree, or strongly disagree?*

The instructions must also make clear when it is acceptable for an interviewer to expand on a question, including paraphrasing the question or supplying a definition. Typically this is not allowed, but there are exceptions, especially in so-called conversational interviewing. For example, if a respondent does not understand a factual question but explains an answer to the interviewer, the interviewer can "work with" the respondent to come up with the correct answer. If, for example, the question is *How many brothers and sisters do you have?* and the interviewer has been supplied with guidelines (such as how to count stepbrothers and stepsisters and half brothers and half sisters), then a conversation between the two is appropriate.

Two important aspects of interviewer training, particularly for volunteer interviewers, relate to confidentiality and neutrality. Confidentiality means that everything that occurs in the interview is confidential. The interviewer should never discuss a respondent's answers with others except project personnel. In addition, the interviewer should never reveal who has been interviewed. This is particularly an issue for community or school surveys where volunteer interviewers may personally know some of the respondents. Indeed, if an interviewer is given a name, address, or telephone number that is familiar, he or she should return that information to a supervisor and not attempt an interview.

The interviewer must also maintain a position of neutrality. That is, respondents should not be able to tell where interviewers stand on issues they are asking about. This can be a problem particularly for volunteer or untrained interviewers. Interviewers should be neutral, not only in their words, but also in their nonverbal behavior and dress. An interviewer who personally disapproves of a respondent's answers may communicate that feeling through raised eyebrows, a frown, or tone of voice. Such nonneutral behaviors can significantly bias subsequent responses.

While neutrality is a must, interviewers should be supplied with information to assist respondents after the interview if circumstances warrant that. For example, if a survey of smokers is about smoking cessation, interviewers should have a telephone number available for quit smoking help. The survey sponsor may decide to offer the information only to respondents who ask for it or to all respondents at the end of the interview. Information on help with domestic violence should be available if relevant to the topic of the questionnaire and if the respondent has mentioned such violence. In cases like these, the assistance should be offered to all respondents who mention the problem, whether they request it or not, but the offer should come only after the interview has been completed, so as not to bias subsequent questions.

Interviewer Monitoring, Supervision, and Validation

Interviewer training and supervision do not end with the training held prior to the start of data collection. It is important to continue to supervise, including monitoring and validating their work and re-training when necessary. Interviewing that is conducted from a centralized facility offers the ability for greater supervision and monitoring. Telephone calls may be monitored and, with CATI surveys, monitoring can include following along with the questionnaire as viewed on another computer screen as the interview is conducted. With the appropriate computer connections, it is possible for a sponsor or client to monitor interviews from another site. Interviews conducted at other sites (e.g., interviewers calling from their home phones, face-to-face interviews conducted in homes, malls, or clinics) offer fewer opportunities to monitor and supervise, making training and validation even more important.

Validation is the term survey researchers use for checking interviews that have been conducted—to verify to the extent possible that they have actually been conducted, conducted correctly, and with the appropriate respondents. The major way to validate an interview is to recontact the respondent (usually by telephone, regardless of mode of interview), explain the contact as checking the quality of the interview, and ask a few of the questions from the original interview, to verify the integrity of the data. It is standard practice to validate a percentage of each interviewer's work. Sometimes this percentage is greater for less experienced interviewers or for all interviewers at the beginning of a survey. However, it is recommended that validation continue for all interviewers and throughout the entire data collection period. If falsification of data is likely, it may be necessary to validate 100% of that interviewer's cases, both interviews and noninterviews. For CATI interviews, supervisors can also check the computerized time clock, to verify that the interview took as long as it should have (e.g., not 3 minutes for a standard 20-minute questionnaire). If the validity of the data is in doubt, the case(s) should be deleted from the final data set.

Calling Rules

Calling rules (or contact rules) refer to the rules of data collection set up by the survey administrator. They include the dates to begin and end the data collection, the days of the week and times of day that calling will be allowed, how many contacts each potential case should be tried (typically a maximum number), and how to code the outcome of each contact attempt. Except in telephone surveys where a predictive dialer is used to assign an already connected call to an interviewer, interviewers can play an integral part in the success of the study by reviewing the history of each case and deciding the best way to proceed before the next contact is attempted. An important part of this process is to gather information from a prospective respondent or household about the best day of the week and time of day to attempt another contact. The more contact attempts made, the more likely a potential respondent will turn into an actual respondent.

Diane O'Rourke

See also Audio Computer-Assisted Self-Interviewing (ACASI); Bias; Bilingual Interviewing; Calling Rules; Closed-Ended Question; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Confidentiality; Conversational Interviewing; Face-to-Face Interviewing; Fallback Statements; Falsification; Field Period; Field Work; Interviewer; Interviewer Characteristics; Interviewer Effects; Interviewer Monitoring; Interviewer Neutrality; Interviewer Productivity; Interviewer-Related Error; Interviewer Training; Interviewer Variance; Introduction; Measurement Error; Open-Ended Question; Paper-and-Pencil Interviewing (PAPI); Predictive Dialing; Respondent; Respondent–Interviewer Rapport; Standard Definitions; Training Packet; Validation; Verification

Further Readings

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., & Waksberg, J. (1988). *Telephone survey methodology*. New York: Wiley.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.

Maynard, D. W., Houtkoop-Steenstra, H., Schaeffer, N. C., & van der Zouwen, J. (2002). *Standardization and tacit knowledge: Interaction and practice in the survey interview*. New York: Wiley.

Intracluster Homogeneity

Intracluster (or intraclass) homogeneity is a concept related to the degree of similarity between elements in the same cluster. The intracluster (or intraclass) correlation coefficient, ρ, measures the degree of homogeneity among population elements within the sampling clusters. Intracluster homogeneity is computed as the Pearson correlation coefficient between pairs of elements that are in the same cluster.

In terms of the variance components in an analysis of variance (ANOVA), intracluster homogeneity measures the extent to which the total element variance in the population is due to the *between-cluster variance*. In other words, ρ measures intracluster homogeneity in terms of the portion of the total variance that is attributable to cluster membership. When there is complete homogeneity within clusters, the between-cluster variance accounts for all the variance in the population and ρ is equal to 1.0. When there is complete heterogeneity within clusters, the within-cluster variance accounts for all the variance in the population and ρ is a negative number equal to the inverse of the size of the cluster minus 1.0. Finally, when the clusters are comprised on random elements from the population with no relationship to each other, ρ is zero.

In practice, the intracluster correlation coefficient typically is positive, but usually not very close to 1.0. This implies that there is some homogeneity within clusters, with elements from the same cluster being more similar to each other than elements selected at random from the population. In these cases, cluster sampling is less efficient than simple random sampling, necessitating some other gain to cluster sampling, like cost savings, to justify the efficiency loss for a cluster sample.

Cluster sampling is frequently used in practice because often it is not feasible or possible to compile sampling frames that consist of all population elements, especially when sampling large human populations. In addition, many times the costs of face-to-face interview data collection are prohibitive when sampling large human populations that are geographically dispersed.

For example, a complete sampling frame of all K–12 public school students in the United States does not exist, and it would be prohibitively expensive for any survey organization to construct such a sample frame. On the other hand, a complete frame of all K–12 public schools in the United States may be available from various sources, and a complete frame of students within each school is usually available. Therefore, a sample of students may be selected in two stages. In the first stage, a sample of schools is selected from the frame of all schools. In the second stage, a sample of students is selected from the frame of all students within each selected school.

Under this sample design, each school constitutes a sampling cluster, and the final sample consists of all sampled students from all sampled schools. This two-stage cluster sample design may be expanded to incorporate additional sampling stages. For example, one possible four-stage design is to select school districts in the first stage, schools in the second stage, classrooms in the third stage, and students in the fourth and final stage. Thus, cluster sampling allows the sample to be selected in successive stages. The sampling frame at each stage is either readily available or can be conveniently constructed.

Although cluster sampling can lead to considerable cost savings, it also is known to significantly increase the variance of many survey estimates as a result of intracluster homogeneity. Sampling clusters are usually defined by geographic or spatial proximity; that is, population elements within the same clusters are close to each other physically. Examples of clusters that are often used in sample surveys include counties, city blocks, census tracts, schools, hospitals, households, and so on. With respect to many population characteristics (demographic, socioeconomic, political, behavioral, epidemiological, health care, and the like), elements in the same clusters tend to be more similar than those in different clusters, resulting in a positive correlation among elements in the same clusters. By confining the sample to a subset of clusters, cluster sampling tends to reduce the spread and representativeness of the sample. Compared to a simple random sample of the same size (in terms of the number of elements), a cluster sample is more likely to lead to extreme estimates and hence increased sampling variance.

Intracluster homogeneity is an important tool for measuring sample efficiency and for survey planning. The efficiency of a complex sample design may be measured by the *design effect* (*deff*), defined as the ratio of the variance under the complex design to the variance of a simple random sample of the same size. For a complex design that involves unequal selection probabilities, stratification, and clustering, the design effect may be decomposed into three multiplicative components: (a) weighting effect, (b) stratification effect, and (c) clustering effect. The clustering effect of a two-stage cluster sample may be expressed as

$$deff = 1 + (m - 1)\rho,$$

where m is the size of the subsample selected from each cluster. When the subsample size differs across clusters, the average subsample size may be used as an approximation. The clustering effect of a three-stage cluster sample may be approximated by

$$deff = 1 + (m_1 - 1)m_2\rho_1 + (m_2 - 1)\rho_2,$$

where m_1 is the average number of secondary sampling units (SSUs) selected within each primary sampling unit (PSU), m_2 is the average number of ultimate sampling units selected from each SSU, ρ_1 is the intracluster homogeneity between SSUs within the same PSU, and ρ_2 is the intracluster homogeneity between the population units within the same SSU. An estimated design effect can then be used to determine the sample size, effective sample size, and other relevant design parameters given the variance requirements and/or cost constraints of the survey.

Y. Michael Yang

See also Analysis of Variance (ANOVA); Clustering; Cluster Sample; Design Effects (*deff*); Effective Sample Size; Primary Sampling Unit (PSU); ρ (Rho); Sampling Variance

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Kish, L. (1987). *Statistical design for research*. New York: Wiley.

INTRODUCTION

The survey introduction is a key step that affects the survey's response rate; therefore, interviewers need special training in it. A key part of an interviewer's role is to gain cooperation from the respondent. This opportunity for enlisting cooperation occurs within a time period of variable length, which starts with the interviewer's initial contact with the sampled unit and continues until the selected respondent agrees to participate in the survey or provides a definitive "no." Depending on the survey design, this conversation is often conducted over multiple callbacks. This time period has traditionally been called the *doorstep introduction*. In this entry, it is referred to as the *survey introduction*.

Interviewers' abilities to gain cooperation during the introduction to a survey vary greatly. For example, in exploring interviewers' survey introduction skill versus respondents' reluctance, Pamela Campanelli and her colleagues found that interviewers' skill could affect response rates by 13 to 20 percentage points for in-person surveys.

This entry details the five parts of the survey introduction and discusses the differences and similarities of introductions for surveys of establishments and surveys of households.

Parts of Introduction

The survey introduction can be thought as being made up of five parts, which generally occur in the following order.

1. Interviewers Introduce Themselves

In this first step the interviewers are typically trained to introduce themselves by saying their name; the name of the survey organization, sponsor, or both; and the topic of the survey. For example,

> Hello, I'm Pamela Jones calling from Research Inc. We are doing a study about health care and doctors for the ABC Institute of Health.

Although this sounds simple, this is a very critical step. As suggested by the University of Michigan's *Interviewer's Manual*, interviewers must convince the respondent that he or she is a professional interviewer from a reputable organization, who is collecting valuable and important data and that the respondent is key to the success of the research.

Interviewers' voice, manner, and beliefs about themselves make as much of a difference as what they say. Interviewers should show a warm, friendly, confident manner and speak slowly. If interviewers believe they will have difficulty and are not confident, this will show.

For in-person surveys, the interviewer's appearance can also be a factor in providing credibility. At the U.K. National Centre for Social Research, interviewers are not told specifically what to wear, but they are told to be "neat and neutral." For in-person surveys, interviewers have the advantage of being able to show their identification card, and it is generally suggested that interviewers smile and make eye contact.

2. Advanced Letters

Use of an advance letter is common for in-person surveys and also can be used in many telephone surveys. An advance letter gives interviewers a psychological advantage. They are not making a cold call but one that has been warmed by the legitimacy of the advance letter. Even if the person who answers the door (or answers the phone) has not received or read the letter, interviewers can still use the letter to their advantage. Well-trained interviewers have copies of the letter with them (or at least know the content) and use this as a peg to start a conversation:

> Have you read our letter? . . . No? . . . This is what it looks like. As you can see, it shows you that . . .

Well-trained in-person interviewers also carry other relevant materials that will help demonstrate the credibility and usefulness of the survey, such as newspaper clippings showing the informative results of past surveys.

3. Finding the Right Person

This step is about within-household selection and screening. If the goal of the survey is to interview one person per household, this needs to be done systematically (ideally randomly) to avoid the bias connected with simply interviewing the most cooperative or available person who answers the door or the phone.

Random procedures differ for in-person and telephone surveys. For an in-person survey, the interviewer may be trained to use an adaptation of the classic procedure designed by Leslie Kish. As outlined in the University of Michigan's *Interviewer's Manual*, this procedure is as follows:

> In order to determine whom I need to interview, I need to know who lives at this address. *[Each household member's name and age are obtained and their relationship to the informant.]*
>
> You have said the household members are *(REPEAT NAME LISTING)*; does that include everyone living here at the present time?
>
> Now I will use a selection procedure—I am going to number the persons in this household to determine whom we need to interview—it will take a second . . .

This list of numbered household members is then compared to a random selection table that can either be pre-printed on a paper questionnaire (or attached via a label) or computer-determined for a computer-assisted questionnaire.

In the case of a telephone survey, many organizations feel this is much too cumbersome, especially because the initial few minutes of contact are even more critical as it is much easier for the respondent to break off contact with the interviewer. Popular quasi-random alternatives are the use of the last or next birthday method. The last birthday method would proceed as follows:

> To get a random sample, I need to speak with the adult in the household who had the most recent birthday.

Although early studies suggested the birthday methods produced equivalent data to the Kish approach, more recent studies, such as one by Cecilie Gaziano, point out that the birthday procedures do not necessarily work as accurately as desired.

Other within-household selection issues occur when the survey is designed for a special subgroup of the population and there is no list of the members of this subgroup. In this case, households are contacted as per Step 1, and then a short, screening questionnaire is administered by the interviewer to determine if any one in the household is a member of the special subgroup. For example, a sample of elderly can be selected from a sample of households. This is more problematic in a telephone survey, as screening questionnaires can contain sensitive and intrusive questions and this can interfere with initial rapport being achieved by the interviewer.

In the case where selection or screening is needed, it is likely that the person who answers the door or phone will not be the person who is finally selected. Nonetheless, it is extremely important for interviewers to build good rapport with this household informant so that the needed information can be obtained. Also, if the informant becomes convinced of the benefits of the survey, he or she can be a good influence on the household member who has been chosen as the respondent. It should be noted that in some cases, the informant might say that the selected respondent will definitely refuse. In such cases, interviewers need to be diplomatic and make it clear that although the informant may be right, their job requires that they talk with the selected person. In some cases, the informant (serving as a "gatekeeper") may deny access to the selected respondent. Gaining access may require sensitivity and perhaps use of a different interviewer.

Not all surveys require selection or screening. It could be that all the members of a household are to be interviewed. Or in some cases, only a knowledgeable adult is sought to provide household level information. In rarer cases, proxy response is allowed with a knowledgeable adult answering for themselves as well as other household members. But although such a design saves time and money, there is a trade-off with quality. Proxy reporting can lead to poorer-quality data.

4. Handling Respondent Questions and Concerns

At any point in Steps 1, 2, and 3, the respondent may interrupt with questions, such as "How long will it take?" "How did you choose me?" "What's it all about?" Or negative statements, such as "I'm too busy" or "I'm not interested."

This is the time in the introduction that the interviewer's skill is most needed. To be successful, interviewers need to be adept in the following ways.

Be Prepared

Interviewers who are prepared will increase credibility and those who are not prepared will lose credibility. Take the following survey introduction excerpt recorded by Campanelli and her colleagues from an actual interview.

Respondent: What's the research about?

Interviewer: It's called political tracking. If you come to any question you're not sure of or don't know how to answer you just leave it, put don't know, things like that.

Respondent: Right, so it's about politics?

Interviewer: Well just how you think, what you think, that sort of thing.

In this example, it does not appear the interviewer knew anything about this survey. Instead, interviewers need to know fully about the study, they need to prepare a variety of accurate and courteous explanations in their own words. And they need to practice these outside of the interview situation. Even the best clarification can be damaged by a nervous-sounding voice.

Have the Right Mind-Set

The right mind-set makes a big difference during the introduction. Creating this is something that the interviewer can prepare in advance. For example, the University of Michigan's *Interviewer's Manual* suggests that interviewers should assume the following:

- It is the perfect time to do the interview.
- The respondent will be friendly and interested.
- If the respondent isn't interested, it is because he or she doesn't know fully about the survey yet.

Eleanor Singer and her colleagues found that interviewers' expectations about the ease of persuading respondents to be interviewed were significantly related to their response rates.

Use Active Listening/Observation

Everyone appreciates being listened to. Accurate listening to the respondent is vital during the survey introduction, but ironically it is most challenging because of the pressure to secure an interview. Good interviewers need to listen not only to the words respondents are saying but also to their tone of voice and the deeper meaning heard through their words. In-person interviewers have the added advantage of seeing respondents' facial and body language. It is also important to listen for sounds in the respondent's environment, which give clues to the respondent's situation, for example, a crying infant.

Use Tailoring

Every respondent is an individual, so general standardized replies to respondent concerns are the least effective. Robert M. Groves and Mick P. Couper have found that to be successful, interviewers need to remain flexible and be fully prepared to tailor their manner and explanation for each and every situation. It also is recommended that interviewers analyze how their reply is working and if it isn't proving successful, to try a different one.

Be Succinct

Good interviewers make their replies clear, coherent, and to the point—thus, providing long rationales should be avoided. Good interviewers should also acknowledge the respondent's viewpoint and never argue directly with the respondent during the introduction (or at any time, for that matter).

Maintain Interaction

Groves and Couper demonstrate the importance of the interviewer maintaining interaction with the respondent. The key here is that this allows rapport to develop, but more importantly it allows the interviewer to gain more information about the respondent's true concerns so that these can be addressed. Campanelli and her colleagues found that some interviewers were very good at shifting the interaction from "official" to "personal" so that the respondent is seen as an individual. But this needs to be done with sincerity and without being overly familiar.

Retreat and Reapproach

Respondent reluctance is often specific to a particular situation. Good interviewers need to watch for signs and back off before receiving a final "no." It is much better to retreat and leave the door open for another contact than to get an outright refusal.

Don't Ask Questions That Can Be Answered "No"

Good interviewers are careful not to ask questions in the introduction that can lead to the respondent's easily answering "no." For example,

> *Problematic response:* I've called at a bad time. Is it okay if I call back?
>
> *Better response:* I've called at a bad time. I'll call back another time.

5. Tactics for Subsequent Callbacks

On subsequent callbacks, interviewers ideally will reflect back to the respondent the issues that the respondent had mentioned previously. For example:

> I know that you are very busy. I was hoping that your schedule might have freed up a bit.
>
> You mentioned that you were busy getting ready for your daughter's wedding. I was hoping that now would be a better time for you.

This is to show that the interviewer has heard what the respondent said before. Then the subsequent call does not come across as bothering the respondent. Rather, it shows the respondent that he or she is very important to the success of the study.

Survey Introductions for Surveys of Establishments

When conducting surveys of establishments, there are several similarities and differences to the survey introduction. As opposed to household surveys, finding the right person to interview takes on new difficulties. Prior to the survey, one must precisely define the organizational entity that one is interested in, the location of the business unit of interest, and the name of the person within the company who will be appropriate to complete the survey. Don Dillman discusses how surveys addressed to the company can have response rates 30 percentage points lower than surveys directed to a named individual within the company. Exploratory telephone calls are often used to find out this key information. Once this is known, the respondent may be contacted by telephone with an invitation to participate, and this would follow closely the steps for household surveys (outlined earlier in this entry). In other cases, a self-administered mode may be used where the respondent is sent a postal questionnaire (and may be given a Web option) and telephone contact is limited to a follow-up method after several reminders are sent.

Pamela Campanelli

See also Advance Contact; Designated Respondent; Gatekeeper; Household Informant; Interviewer Neutrality; Refusal Avoidance; Respondent–Interviewer Rapport; Screening; Tailoring; Within-Unit Selection

Further Readings

Campanelli, P., Sturgis, P., & Purdon, S. (1997). *Can you hear me knocking: An investigation into the impact of interviewers on response rates*. London: National Centre for Social Research.

Dillman, D. (2000). *Mail and Internet surveys*. New York: Wiley.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*(1), 124–157.

Groves, R., & Couper, M. (1998), *Nonresponse in household interview surveys*. New York: Wiley.

Kish, L. (1949). A procedure for objective respondent selection within a household. *Journal of the American Statistical Association*, *44*, 380–387.

Morton-Williams, J. (1993). *Interviewer approaches*. Aldershot, UK: Dartmouth Publishing.

Singer, E., Frankel, M., & Glassman, M. (1983). The effect of interviewer characteristics and expectations on response. *Public Opinion Quarterly*, *47*, 68–83.

Survey Research Center. (2001). *Interviewer's manual*. Ann Arbor: University of Michigan, Institute for Social Research.

INVERSE SAMPLING

Inverse sampling is an adaptive sampling technique credited to J. B. S. Haldane's work in the 1940s. Under many study designs, it is desirable to estimate the frequencies of an attribute in a series of populations, each of which is much larger than the sample taken from it so that the population size is assumed to be infinite. However, the probability of the attribute occurring in some of the populations may be so small that under a fixed sample size design, not enough cases of interest are selected to estimate the attribute of interest.

Inverse sampling draws from the negative binomial distribution in that a series of Bernoulli trials are conducted until a predefined r number of successful cases occur. Usually, r is the desired number of cases from the population with the smallest probability of selecting a case with the attribute of interest. Under this design, the total sample size is a random variable. Therefore, traditional estimates, based on the binomial distribution, of an attribute's probability of occurrence are biased. However, it can be shown that if the total sample size is X, then the uniformly minimum variance unbiased estimator for the probability p is $\hat{p} = (r-1)/(X-1)$. However, D. J. Best derived the variance for this estimator and showed that it is intractable as a function of p or of r. Therefore, only an upper bound, such as the one proposed by Govind Prasad and Ashok Sahai, can be derived for the variance.

Applications

Applications for inverse sampling can have broad appeal. One such application is the ability to determine the better of two binomial populations (or the one with the highest probability of success). For example, in a drug trial, where the outcome is success or failure, inverse sampling can be used to determine the better of the two treatment options and has been shown to be as efficient, and potentially less costly, than a fixed sample size design. Milton Sobel and George Weiss present two inverse sampling techniques to conduct such an analysis: (1) *vector-at-a-time* (VT) sampling and (2) *play-the-winner* (PW) sampling. VT inverse sampling involves two observations, one from each population, that are drawn simultaneously. Sampling continues until r successful observations are drawn from one of the populations. PW inverse sampling occurs when one of the populations is randomly selected and an observation is randomly selected from that population. Observations continue to be selected from that population until a failure occurs, at which point sampling is conducted from the other population. Sampling continues to switch back and forth between populations until r successful observations are selected in one of the populations. Under both VT and PW, the population from which r successes are first observed is determined the better population. In clinical trials PW is an advantageous design because it has the same probability requirements as VT, but the expected number of trials on the poorer population is always smaller. Sobel and Weiss have also extended this methodology for $k \geq 2$ populations.

Inverse sampling is also used to estimate the number of events that occur in an area of interest based on a Poisson distribution. In these situations, one can use inverse sampling to estimate the total number of events or the number of events during a certain period by selecting a sampling unit and counting the number of events that occur in that unit. A series of independent units are sequentially selected until the total number of events across all of the selected units meets or exceeds a pre-assigned number of events. The number of trials needed to reach the pre-assigned number of events is then used to estimate the mean number of events that will occur. This design assumes a Poisson distribution but not a Poisson process. Therefore, not every sampling unit selected has to have a Poisson distribution, but all the sampling units combined do have a Poisson distribution. An example of this design would be to estimate the number of accidents on the road. Because the number of accidents depends on the day of the week, a week would be the smallest sampling unit that one could assume had a Poisson distribution. If one day were the sampling unit, then a Poisson distribution might not always hold.

Inverse Sampling Designs

For finite population sampling, two inverse sampling designs have been developed over the past several years: *multiple inverse sampling* (MIS) and *general inverse sampling* (GIS). Under an MIS design, the population is partitioned into two or more subpopulations with known sizes. MIS is effective when one of these subpopulations is rare and it would be undesirable to obtain none or very few responses from the rare subpopulation. MIS sequentially selects sampling units, without replacement, until the predetermined sample sizes are obtained for all subpopulations. Sampling from a subpopulation is ceased once its quota is met. As with inverse sampling, the total sample size under MIS is a random variable. Therefore, Horvitz-Thompson estimators for finite populations are biased. As an alternative, two unbiased means and variances have been proposed. Kuang-Chao Chang, Jeng-Fu Liu, and Chien-Pai Han have derived formulas to compute the expected mean and variance of the final random sample size as well as an unbiased estimator and its variance. Also, Mohammad Salehi and George Seber showed that Murthy's estimator can be used to construct unbiased estimators of the mean and variance in a sequential sampling design. A situation where MIS is applicable is the estimation of a species of bird's prevalence in various geographic regions where the total population in the region is known and the species is thought to be rare in at least one of the regions. Sequential sampling of all birds occurs until the predetermined number of the species is selected in each region. Once a region has reached its quota, sampling from that region is ceased.

In MIS, when an attribute is very rare, it is possible that the expected total sample size is either logistically too large to achieve or results in a census. GIS addresses this issue. Like MIS, GIS is an adaptive sampling procedure where one divides the population into predefined subpopulations. Then, a preliminary simple random sample is drawn across all subpopulations. Sampling is completed if the initial sample contains a prespecified number of units in each subpopulation cell. Otherwise, a sequential sample is drawn until either the prespecified number of units within each subpopulation cell is met or the total sample size reaches a predetermined amount. By including this second condition, GIS limits the total sample and, therefore, prevents the total sample size from being unmanageable. GIS uses Murthy's estimator and its variance to estimate population parameters. Using the MIS example, under GIS, a simple random sample of birds would be drawn from all regions. If a pre-set number of the species of interest was selected from the region where the bird is most rare then sampling would end. Otherwise, sequential sampling of birds would continue until the predetermined number of this species of bird was selected or a predetermined sample size was selected.

Marcus Berzofsky

See also Adaptive Sampling; Sequential Sampling

Further Readings

Chang, K.-C., Liu, J.-F., & Chien-Pai, H. (1998). Multiple inverse sampling in post-stratification. *Journal of Statistical Planning and Inference*, *69*, 209–227.

Haldane, J. B. S. (1945). On a method for estimating frequencies. *Biometrika*, *33*(3), 222–225.

Salehi, M. M., & Seber, G. A. F. (2004). A general inverse sampling scheme and its application to adaptive cluster sampling. *Australian & New Zealand Journal of Statistics*, *46*(3), 483–494.

Sobel, M., & Weiss G. H. (1971). Play-the-winner rule and inverse sampling in selecting the better of two binomial populations. *Journal of the American Statistical Association*, *66*(335), 545–551.

Weiler, H. (1972). Inverse sampling of a Poisson distribution. *Biometrics*, *28*(4), 959–970.

Issue Definition (Framing)

Issue definition, or framing, refers to the careful use of language or other symbols in public discourse. If individual frames reside in both mental structures of the mind and in political discourse, framing is a social process that links the two. As such, framing is an important construct in the measurement and understanding of public opinion. Elected officials, other key decision makers, special interest groups, journalists, scholars, lobbyists, and pollsters are among those interested in issue definition (and its measurement) because of its importance in public policy formation and acceptance in modern democracies.

Public issues are inherently matters about which people disagree. There are often powerful financial interests backing one side or another. Other times, profound moral principles may be at stake. This accounts for the importance of public issues and the difficulties in studying them. Language used to describe political issues is not static. As the major political parties and interest groups fight over issues, the key battleground is political language. Interest groups struggle to gain acceptance for their terms and ideas and to have them incorporated into political dialogue. To win the battle over the acceptance of key terms is often the key to a wider political victory.

One key aspect of issue framing is that the terms used are carefully chosen by political and other intellectual elites to convey messages that resonate in particular ways with key elements of the public. As Donald Kinder and Lynne Sanders have noted, frames lead a double life. What they mean is that issue frames are both literary devices and mental devices. Frames are powerful because they are precisely chosen terms or phrases that resonate with the ways humans are programmed to think.

The most successful frames for issues are never casually chosen. They are carefully cultivated by elite communicators—such as politicians, political pundits, and prominent media correspondents—who know exactly what they are doing. Sociologist William A. Gamson is a leading figure in studying how public issues are described in the mass media, and how people learn these ideas, use them, and reproduce them in their daily lives. Gamson's research points to a series of framing devices and reasoning devices that together make up an "issue package." Issues that are in front of the public for a long period of time will eventually have fully developed frame structures that are quite complete. Particularly complex issues will also have a variety of competing frames associated with them. One of the key framing devices is *metaphor*. When a political figure uses a metaphor such as "war" to describe the appropriate response to a social problem such as terrorism, this choice is made advisedly. The term is chosen in such a way that alternative ways of thinking about the problem do not come to mind. Effective framing often is done in a way that seems so natural and uncontroversial that other ways of conceptualizing the issue can hardly be imagined.

The effective framing of a problem is often inclusive of preferred solutions to the problem. For example, the George W. Bush administration's choice of phrases for framing the global struggle with terrorism in the wake of attack and destruction of the World Trade Center as the "War on Terrorism" emphasized the use of organized military force to maintain public order and allay the fears of the population. The war metaphor also diverted thinking away from alternative forms of force such as police or detective work, as well as various soft-power strategies of public diplomacy and cultural engagement.

Once frames are established in mainstream public discourse, they become conventional, and journalists and others start to use them in routine ways. This tends to close off alternative interpretations and gives the impression that other points of view have been discredited or no longer apply. When a particular way to frame something is ascendant, the focus tends to be diverted away from an examination of the nature of the highly charged political phrase and what political advantage it is providing.

The challenge for pollsters and other survey researchers is to regularly use the most neutral terms possible but still convey accurately what public policy issue is being studied. This can be difficult, particularly when phrases enter popular discourse in an uncritical manner. Many academic surveys contain embedded experiments to examine the effects of question wording on survey response. Simple wording variations involve the order of names or whether descriptions or additional labels are provided to respondents. These investigations can be viewed as *framing effect studies* and may have substantive interpretations in public opinion research.

Depending on the language chosen, surveys and polls can measure and analyze public opinion but also sometimes help manufacture a particular response. This is, on one hand, a problem of question wording variation and the kind of random noise that can affect survey questions. But there is also the possibility of using aspects of real public discourse in the framing of questions. If survey research were to adapt question language that mimics actual public discourse by political elites, it would open the door to substantive interpretations of question wording effects. However, an advantage of this approach would be to help understand the dynamic role of language in political discourse and in the formation of public opinion. The literature of political journalism and the framing activities of social movements provide many examples to help formulate such studies and interpret the results.

Gerald M. Kosicki

See also Public Opinion; Public Opinion Research; Questionnaire-Related Error; Question Wording as Discourse Indicators

Further Readings

Kinder, D. R. (2003). Communication and politics in the age of information. In D. O. Sears, L. Huddy, & R. Jervis (Eds.), *Oxford handbook of political psychology* (pp. 357–393). New York: Oxford University Press.

Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate with survey questions: The case of white opinion on affirmative action for blacks. *Social Cognition*, *8*, 73–103.

Lakoff, G. (2002). *Moral politics: How liberals and conservatives think*. Chicago: University of Chicago Press.

Lakoff, G. (2004). *Don't think of an elephant!* White River Junction, VT: Chelsea Green.
Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication*, *10*, 55–75.
Rochford, D. A., & Cobb, R. W. (1994). *The politics of problem definition*. Lawrence: University Press of Kansas.

Item Order Randomization

Most social scientists are aware that responses to survey questions can be significantly affected not only by how questions are worded but also by the order in which the questions are asked. Furthermore, they are generally aware that the order in which the response alternatives within a question are presented can likewise have a significant influence on survey results. Despite this awareness of order effects in surveys, many investigators either ignore these potential sources of measurement error in designing their questionnaires or fail to systematically control for them by fully randomizing the order in which the items are presented.

Most researchers who suspect there is the potential for an order effect in the questionnaire they are designing will rotate the items, typically presenting them in order X on one form of the questionnaire and order Y on the other. A prototypical example of this practice comes from a November 1997 Gallup poll. On one form of the questionnaire, respondents were asked the following questions in this sequence:

1. *How likely is it, in your view, that a terrorist group will attack the United States using* chemical or biological weaponry *sometime within the next ten years—very likely, somewhat likely, somewhat unlikely, or very unlikely?*
2. *How likely is it, in your view, that a terrorist group will attack the United States using* nuclear weaponry *sometime within the next ten years—very likely, somewhat likely, somewhat unlikely, or very unlikely?*

On the other form of the questionnaire, respondents were asked the same questions in reverse sequence. These alternative sequences produced a significant question order effect. In this case, where there are only two questions to consider, simply reversing the sequence of the items amounts to fully randomizing them and controlling for order effects. But if there are three or more related questions to be asked in a sequence, then full randomization requires more than just asking the questions in one order and the reverse order. With three questions to be asked, for example, there are $3 \times 2 \times 1$, or six possible permutations of the order in which the items can be presented: Q1-Q2-Q3, Q1-Q3-Q2, Q2-Q1-Q3, Q2-Q3-Q1, Q3-Q1-Q2, and Q3-Q2-Q1. It is relatively rare, however, to see such a fully randomized order of item presentation in a survey questionnaire, particularly when the data must be collected using a noncomputerized, paper-and-pencil questionnaire, because of the cost and the impracticality of administering six separate versions of the questionnaire. But fully randomized designs do not appear to be that much more common in computerized telephone and personal interviewing or in Web-based surveys, in which they can be readily implemented by programmers.

Aside from the additional staff time it may take to program such designs, researchers may resist doing so because of the small subsample sizes that will be available to analyze the effects of the different orders of item presentation. For a national sample of 1,000 cases, for example, a fully randomized design of six separate conditions will result in approximately 166 respondents per subgroup, which makes it difficult to detect statistically significant effects and to automatically control for nonrandom measurement errors. These must still be analyzed systematically. Given current practices of simple question or response alternative rotation, many potential order effects in existing data sets have yet to be discovered and controlled for through item order randomization, or even properly understood.

George F. Bishop

See also Context Effect; Measurement Error; Question Order Effects; Random Assignment; Random Order; Random Start; Response Order Effects; Split-Half

Further Readings

Bishop, G. F. (2005). *The illusion of public opinion: Fact and artifact in American public opinion polls*. Lanham, MD: Rowman & Littlefield.
Schwarz, N., & Sudman, S. (1992). *Context effects in social and psychological research*. New York: Springer.
Smith, T. W. (1983). An experimental comparison of clustered and scattered scale items. *Social Psychology Quarterly*, *46*, 163–168.

ITEM RESPONSE THEORY

Item response theory (IRT) is an approach used for survey development, evaluation, and scoring. IRT models describe the relationship between a person's response to a survey question and his or her standing on a latent (i.e., unobservable) construct (e.g., math ability, depression severity, or fatigue level) being measured by multiple survey items. IRT modeling is used to (a) evaluate the psychometric properties of a survey, (b) test for measurement equivalence in responses to surveys administered across diverse populations, (c) link two or more surveys measuring similar domains on a common metric, and (d) develop tailored questionnaires that estimate a person's standing on a construct with the fewest number of questions. This entry discusses IRT model basics, the application of IRT to survey research, and obstacles to the widespread application of IRT.

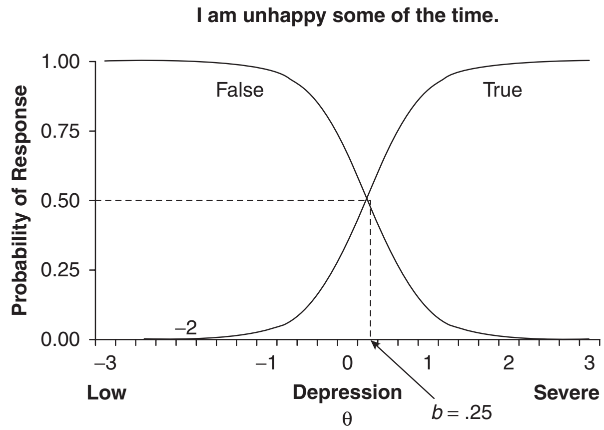

Figure 1 Item response curves representing the probability of a "false" or "true" response to the item *I am unhappy some of the time* conditional on a person's depression level. The threshold ($b = 0.25$) indicates the level of depression g(θ) needed for a person to have a 50% probability for responding "false" or "true."

Note: Numbers on the θ-axis are expressed in standardized units and, for the illustrations in this discussion, the mean depression level of the study population is set at 0 and the standard deviation is set to 1. Thus, a depression score equal to ($\hat{\theta} = 2.0$) indicates that a person is 2 standard deviations above the population mean and is highly depressed.

IRT Model Basics

Item Response Curves

IRT models describe for each item in a scale how the item performs for measuring different levels of the measured construct. For example, the item *I don't seem to care what happens to me* would have IRT properties reflecting it is informative for measuring people with severe levels of depression, and an item such as *I am happy most of the time* would have IRT properties reflecting it is informative for measuring people with low levels of depression.

The probabilistic relationship between a person's response to an item and the latent variable (θ) is expressed by item response curves (also referred to as category response curves or item trace lines). For example, Figure 1 presents the IRT response curves for the item *I am unhappy some of the time*, which has two responses, "false" and "true," and is part of a scale measuring depression.

Individuals with little depression are located on the left side of the θ continuum in Figure 1, and people with severe depression are located on the right side of the axis. The vertical axis in Figure 1 indicates the probability that a person will select one of the item's response categories. Thus, the two response curves in Figure 1 indicate that the probability of responding "false" or "true" to the item *I am unhappy some of the time* depends on the respondent's depression level.

The response curves in Figure 1 are represented by logistic curves that model the probability *P* that a person will respond "true" to this item (*i*) as a function of a

$$P(X_i = true|\theta,a_i,b_i) = \frac{1}{1 + e^{a_i(\theta - b_i)}} \qquad (1)$$

respondent's depression level (θ), the relationship (*a*) of the item to the measured construct, and the severity or threshold (*b*) of the item on the θ scale. In IRT, *a* and *b* are referred to as *item discrimination* and *threshold parameters*, respectively.

The item threshold or severity level (*b*) is the point on the latent scale θ at which a person has a 50% chance of responding "true" to the item. In Figure 1, the item's threshold value is $b = 0.25$, which indicates that people with depression levels a quarter standard deviation above the population mean have a 50% chance of indicating "false" or "true" to the question.

The discrimination or slope parameter (a) in Equation 1 describes the strength of an item's ability to differentiate among people at different levels along the trait continuum. In Figure 1, the slope at the inflection point (i.e., the point at which the slope of the curve changes from continuously increasing to continuously decreasing) is $a = 2.83$. The larger the a parameter is, the steeper the curve is at the inflection point, and steeper slopes indicate that the *item characteristic curve* increases relatively rapidly, such that small changes on the latent variable lead to large changes in item-endorsement probabilities. The a parameter also describes the relationship between the item and the trait being measured by the scale, such that items with larger slope parameters indicate stronger relationships with the latent construct.

IRT Model Information Functions

Another important feature of IRT models is the information function, which describes the extent to which an item is useful for measuring persons at different levels of the underlying latent construct, with higher information denoting more precision. Figure 2 presents the item information functions that are associated with three depression items (respondent's trait level [θ] is shown on the horizontal axis and information magnitude is shown on the vertical axis).

The shape of the item information function is determined by the item parameters. The higher the item's discrimination (a) is, the more peaked the information function will be. Thus, higher discrimination parameters provide more information about individuals whose trait levels (θ) lie near the item's threshold value. The item's threshold parameter(s) (b) determines where the item information function is located. In Figure 2, the item *I don't seem to care what happens to me* is informative for measuring high levels of depression, the item *I am unhappy some of the time* is informative for measuring moderate depression levels, and the item *I cry easily* is not informative for measuring any depression level relative to the other items.

The item information functions are a powerful tool because they allow questionnaire developers to reduce respondent burden or to create shorter questionnaires by selecting the most informative set of items that are relevant for the population under study (e.g., a researcher working with a clinically depressed population could select items best for assessing high levels of depression). Items with low information (low discrimination) function may indicate that this particular item has a problem because (a) the content of the item does not match the construct measured by the other items in the scale, (b) the item is poorly worded and needs to be rewritten, (c) the item is too complex for the educational level of the population, or (d) the placement of the item in the survey is out of context.

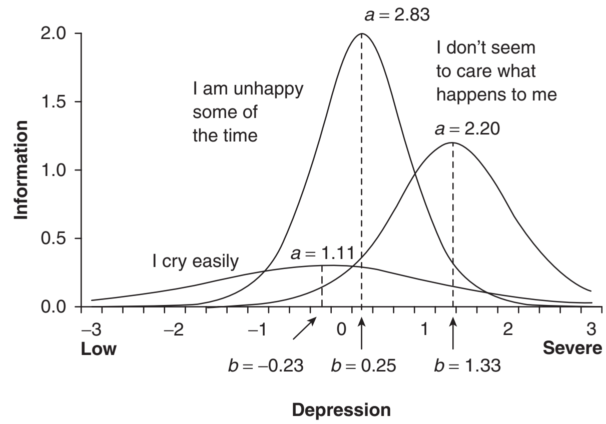

Figure 2 Item information functions for three items: *I cry easily, I am unhappy some of the time*, and *I don't seem to care what happens to me*.

The individual item information functions can be summed across all of the items in a scale to produce the scale information function as shown in Figure 3 (the associated reliability [r] is also provided on the vertical axis). Overall, the scale is highly reliable for measuring moderate to severe levels of depression (i.e., when reliability is above $r = .90$). However, scale precision worsens for measuring persons with low levels of depression.

Family of IRT Models

There are more than 100 varieties of IRT models to handle various data characteristics, such as dichotomous and polytomous response data, ordinal and nominal data, and unidimensional and multidimensional data. The common parametric unidimensional IRT models include the one-, two-, and three-parameter logistic models for dichotomous response data and the graded response model, partial credit model, rating scale model, and generalized-partial credit model for polytomous response data.

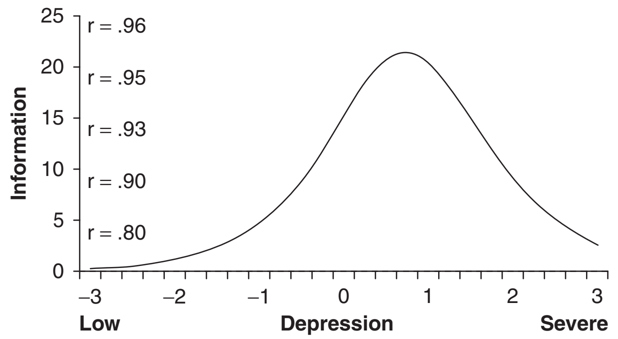

Figure 3 Scale information function for a 57-item depression questionnaire. The horizontal dashed lines indicate the level of reliability (r) associated with different levels of information.

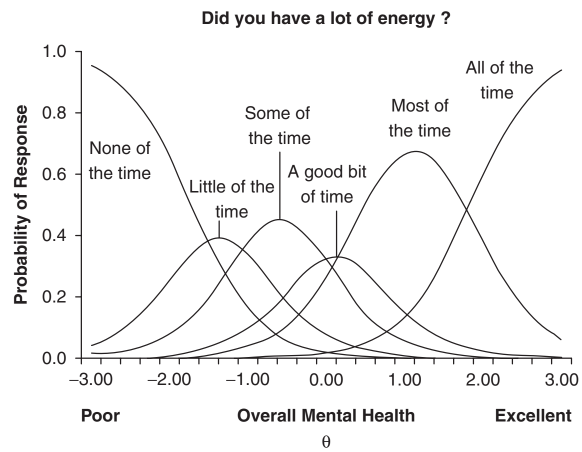

Figure 4 Category response curves representing the probability for selecting one of the six response options for the item, *Did you have a lot of energy?* conditional on a person's mental health level.

Figure 4 presents IRT category response curves (estimated from the graded response model) for a polytomous response question, *Did you have a lot of energy?* which appears in a health status questionnaire. In the figure, there is a curve associated with each of the six possible responses, which models the probability of endorsing the response conditional on a person's level of mental health.

IRT Model Assumptions

The IRT models described in the previous sections make three key assumptions about the data: (1) unidimensionality, (2) local independence, and (3) monotonicity. These assumptions should be evaluated before any IRT model results are interpreted, but IRT models are robust to minor violations of the assumptions, and no real data ever completely meet the assumptions.

The unidimensionality assumption posits that the set of items measures a single continuous latent construct. Unidimensionality can be evaluated by performing an item-level factor analysis to evaluate the factor structure that underlies the observed covariation among item responses. If multidimensionality exists, the investigator may want to consider dividing the scale into subscales, based on both theory and the factor structure provided by the factor analysis, or consider using multidimensional IRT models.

The assumption of local independence means that the only systematic relationship among the items is explained by the conditional relationship with the latent construct. In other words, if the trait level is held constant, there should be no association among the item responses. Violation of this assumption may result in erroneous decisions when selecting items for scale construction. The impact of local dependence can be evaluated by examining how the item parameters and person scores change when one or more of the locally dependent items are dropped.

The assumption of monotonicity means that the probability of endorsing or selecting an item response indicative of better health status should increase as the underlying level of health increases. Approaches for studying monotonicity include examining graphs of item mean scores conditional on *rest-scores* (i.e., total raw scale score minus the item score) or fitting a nonparametric IRT model to the data that yield initial IRT probability curve estimates.

Applications of IRT Modeling in Survey Research

Much of the development and application of IRT modeling has occurred in educational measurement, where IRT is used to help administer and score educational tests like the SAT (Scholastic Assessment Test) and the GRE (Graduate Record Examination). Other disciplines have realized the value of these applications and are learning how to adapt these methods for (a) evaluating the properties of existing scales and guiding survey revisions, (b) determining measurement equivalence across research populations, (c) linking two or more questionnaires on a common metric, and (d) developing item banks for computerized adaptive testing applications.

Evaluating Existing Scales and Guiding Survey Revisions

IRT modeling makes an excellent addition to the psychometrician's toolbox for developing and revising survey questionnaires. The IRT category response curves help questionnaire developers evaluate how well each of the response categories for each item functions for different levels of the measured construct as well as determine whether more or fewer response categories are needed.

The IRT information curves serve as a useful tool for instrument developers to evaluate how well an item or scale functions for measuring all levels of the underlying construct. Developers can use the information curves to weed out uninformative questions or to eliminate redundant items that provide duplicate levels of information across the construct continuum. Effects on precision for removing items from the scale can easily be evaluated with the scale information function. Also, information curves allow developers to tailor their instrument to provide high information (i.e., precision) for measuring their study population. For example, if a developer wants high precision to measure a person at any level of depression (i.e., high information across all levels of the construct continuum), then the information function in Figure 3 suggests adding more items to the scale (or more response options to existing items) that differentiate among people with low depression levels.

Assessing Measurement Equivalence

Items in a survey questionnaire are carefully written to ensure that they are tapping into the same construct no matter which population is responding to the questions. For example, considerable care is taken when an instrument is translated from one language to another. Despite this careful translation process, it may turn out that although the words are the same, the two populations may hold culturally different views of the question. For example, a common finding in depression questionnaires is that Hispanics are more likely to respond positively to a question such as *I feel like crying* than are non-Hispanics, despite controlling for differences between the two populations' depression levels, perhaps because Hispanics are more likely to believe that crying is an acceptable social behavior. This is known as differential item functioning (DIF).

DIF is a condition in which an item functions differently for respondents from one group than for another. In other words, respondents, with similar levels on a latent trait but who belong to different populations, may have a different probability of responding to an item. Questionnaires containing such items may have reduced validity for between-group comparisons because their scores may indicate a variety of attributes other than those the scale is intended to measure.

IRT provides an attractive framework for identifying DIF items. In IRT modeling, item parameters are assumed to be invariant to group membership. Therefore, differences between the curves, estimated separately for each group, indicate that respondents at the same level of the underlying trait, but from different groups, have different probabilities of endorsing the item. More precisely, DIF is said to occur whenever the conditional probability, $P(X)$, of a correct response or endorsement of the item for the same level on the latent variable differs for two groups.

DIF analysis has been used to detect measurement equivalence in item content across cultural groups, males and females, age groups, between two administration modes such as paper-and-pencil versus computer-based questionnaires, and from one language translation of a questionnaire to another. Also, DIF testing can be used for evaluating question ordering effects or question wording effects.

Linking Two or More Scales

It is common in many research settings for several existing instruments to measure the same construct. Combining or comparing results across studies that use different questionnaires in a meta-analytic study is difficult because the questionnaires may have different lengths, different number of response options, and different types of questions with different psychometric properties. IRT modeling provides a solution through its ability to link the item properties from different scales on to a common metric. Several methodologies exist for linking two (or more) instruments. Ideally, one would administer both instruments to a representative sample and then IRT-calibrate (obtain the properties of) the items simultaneously. Alternatively, a set of items that are common to both instruments can be selected as anchors. The anchor items are used to set the metrics to which items not common to both instruments are scaled. Therefore, instruments with a different number or difficulty of items can be linked by responses to a common set of anchor items.

These applications take advantage of a key feature of IRT models, which is the property of *invariance*. If IRT model assumptions are met, item parameters are invariant with respect to the sample of respondents, and respondent scores are invariant with respect to the set of items used in the scale. After the IRT item parameters are estimated (i.e., calibrated), researchers can choose the most salient items to target a person's level of function with the smallest number of items. This method results in different groups receiving different sets of items; however, any given set of items calibrated by the best-fitting IRT model should yield scores that are on a similar metric.

Building Item Banks and Computerized Adaptive Testing

The IRT principle of invariance is the foundation that researchers use to develop computerized adaptive tests (CATs), which yield tailored instruments that estimate a person's level on a construct (e.g., depression) with the fewest number of items. To accomplish this, a CAT has access in its data bank to a large pool of items that have been carefully selected and calibrated by IRT models (called an *item bank*).

Based on a person's response to an initial item, the CAT selects the next most informative item from the item bank to administer to the respondent. After each response, the computer selects and administers from the item bank the next most informative item until a minimal standard error or maximum number of administered items is obtained. The benefits of CAT technology include (a) decreased respondent burden, (b) reduced "floor and ceiling" effects, (c) instant scoring, and (d) widespread availability of this technology on many platforms (e.g., Internet, handheld devices, computer-assisted telephone interviewing).

Limitations

The survey research field has much to gain from IRT methods; however, there are limitations to widespread application. First, many researchers who have been trained in classical test theory statistics may not be as comfortable with the advanced knowledge of measurement theory IRT modeling requires. In addition, the supporting software and literature are not well adapted for researchers outside the field of educational measurement. Another obstacle is that the algorithms in the IRT parameter-estimation process require large sample sizes to provide stable estimates (from 100 for the simplest IRT model to 1,000 or more for more complex models). Despite the conceptual and computational challenges, the many potential practical applications of IRT modeling cannot be ignored.

Bryce B. Reeve

See also Language Translations; Questionnaire Design; Questionnaire-Related Error; Response Alternatives

Further Readings

Bjorner, J. B., Kosinski, M., & Ware, J. E. (2003). Using item response theory to calibrate the Headache Impact Test (HIT^{TM}) to the metric of traditional headache scales. *Quality of Life Research*, *12*, 981–1002.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Mahwah, NJ: Lawrence Erlbaum.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5), S22–S31.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum.

J

JACKKNIFE VARIANCE ESTIMATION

There are two basic approaches to estimation of the variance for survey data: the Taylor linearization method and the resampling method. The resampling method includes the jackknife, balanced repeated replication (Fay's method as a variant), and bootstrap methods. The resampling method calls for creation of many replicate samples ("replicates" in short) taken from the original sample (called also the full or parent sample). Each resampling method uses a unique way of creating the replicates. Each replicate provides a point estimate of the population parameter of interest and the variability among the replicate estimates forms the basis of estimating the variance of the point estimate.

Let θ be the population parameter to be estimated from the sample and let R replicates be created, from which R replicate point estimates, $\hat{\theta}_j$, $j = 1, 2, \ldots, R$ of θ are obtained. Then the jackknife variance estimator is given by $\hat{V}_J = \sum_{j=1}^{R} c_j(\hat{\theta}_j - \hat{\theta})^2$, where c_j are scaling constants to correct the bias and $\hat{\theta}$ is the point estimate based on the full sample.

A special case of the jackknife variance estimator for the sample mean under simple random sampling is a helpful way to understand the idea behind the method. In this case the maximum number of replicates that can be created in a replicate sample is formed by deleting one sample unit at a time—the number of replicates is equal to the sample size n. Then the jackknife variance formula is given by $\hat{V}_J = \{(1-f)(n-1)/n\} \sum_{j=1}^{n} (\hat{\theta}_j - \hat{\theta})^2$ where $f = n/N$ is the finite population correction, N is the size of the finite population under study, $\hat{\theta}_j = \sum_{i \neq j, i=1}^{n} y_i/(n-1)$ is the sample mean of the j-th replicate sample, and $\hat{\theta} = \bar{y} = \sum_{i=1}^{n} y_i/n$, the sample mean for the full sample. Note that in this special case, $c_j = (1-f)(n-1)/n$ for all $j = 1, 2, 3, \ldots, n$. It is not difficult to show that the variance estimator $\hat{V}_J$ is equivalent to the usual variance estimation formula $\hat{V} = \{(1-f)/n\} \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$. Therefore, the jackknife variance estimator is an unbiased variance estimator for this special case.

However, even when the point estimate is a complex one, such as a ratio estimate, the jackknife variance estimator still gives approximately correct variance estimate if the sample size is large (the technical term for this property is *consistency*). The same is true for other point estimates that are defined in a smooth functional form of sample (weighted) totals or averages—survey weights should be used especially when unequal probability sampling has been used. Such point estimates for which the jackknife is consistent include the ratio, regression coefficient, and correlation coefficient, but not the median or, more generally, quantiles.

When the sample design is complex with stratification and clustering (in one or multi-stage), estimation of the variance of a survey estimate is not simple. The usual approach is to assume that the primary sampling units (PSUs) have been selected with replacement within strata, although replacement sampling is seldom used in practice. Then the variance estimator can be written in a form of PSU level aggregates of the variables involved in the definition of the point estimate. This approach easily incorporates the cluster effect on the variance estimate. So it simplifies the variance estimation substantially. However, the price of this simplification is that the variance estimate is generally overestimated as a result of the assumption of replacement sampling of the PSUs. The overestimation is not serious if the PSU level sampling fraction is small, and mild overestimation is generally accepted—this leads to mildly conservative results when the variance estimate is used in statistical inference. However, if the sampling fraction is not small, the overestimation can be substantial, and incorporation of the finite population correction (fpc) may be helpful to reduce the overestimation. (This issue applies not only to the jackknife method but also to the Taylor method.)

Under the with-replacement sampling assumption, replicates are created by deleting one PSU at a time. So if there are R sample PSUs, there will be R replicates (i.e., $R = \sum_{h=1}^{H} n_h$, where H is the number of strata and n_h is the stratum PSU sample size). The variance estimator given at the beginning is applicable in this situation. To compute $\hat{\theta}_j$, the sample weights have to be modified. If replicate j is defined by deleting a PSU in stratum h, then the sample weights of the sample units in the remaining $(n_h - 1)$ PSUs are inflated by the factor of $n_h/(n_h - 1)$, while keeping all other sample weights unchanged. The scaling constant in the variance formula is given by $c_j = (n_h - 1)/n_h$ under the replacement sampling assumption. If, however, the fpc is not negligible, then it can be incorporated through $c_j = (1 - f_h)(n_h - 1)/n_h$, where $(1 - f_h) = (1 - n_h/N_h)$ is the fpc for stratum h and n_h and N_h are stratum PSU sample and population sizes, respectively, if PSUs have been selected by simple random sampling. For the case that PSUs have been selected by unequal probability sampling such as the probability proportional to size sampling, $f_h = \bar{\pi} = \sum_{i=1}^{n_h} \pi_{hi}/n_h$, which is the average inclusion probability of PSUs selected from stratum h, may be used to avoid excessive overestimation. However, the use of the fpc in this way usually results in underestimate of the variance. (Better corrections are available, but they are beyond the scope of this entry.)

Survey weighting involves not only base sample weighting but also various adjustments, such as nonresponse adjustment, post-stratification ratio adjustment, and so on. Mimicking these adjustments on the replicates, the jackknife variance estimator reflects the effect of the weight adjustments on the variance. This is considered one of the strengths of the jackknife variance estimator.

The degrees of freedom of the jackknife variance estimator are determined by $R - H$ including the situation where $H = 1$ (i.e., no stratification). While it is advantageous to have a large number of replicates so as to have a large number of degrees of freedom, having a large number of replicates requires more effort to create replicates and to calculate the variance, as well as more computer space to store replicate weights. So it is ideal to take a good balance between the advantages and disadvantages of having a large number of replicates. If a smaller number of replicates is desired, then sampling units are randomly grouped to form artificial clusters called random groups even when clustering has not been used. Even for a cluster design, PSUs may be combined to reduce the number of replicates. However, such manipulation should be done carefully because it may introduce some bias in the variance estimate.

A special case, which is often employed in practice, is the stratified cluster design with $n_h = 2$ for all strata. In this case, instead of using all possible $2H$ replicates, one can use only H replicates, selecting one replicate from each stratum. It is equivalent algebraically with the full jackknife for a point estimate that is defined as a linear statistic of sample observations, and its degree of freedom is H, which is the same as for the full jackknife. For nonlinear statistics, the equivalence does not hold. However, the two estimators are still close to each other. For this special case, $c_j = 1$ for all H replicates if the fpc is ignored.

In the jackknife variance estimator formula, $\hat{\theta}$ may be replaced by the average of the replicate estimates, $\bar{\hat{\theta}} = \sum_{j=1}^{R} \hat{\theta}_j/R$. However, the difference is usually small, so it does not generally matter which is used. The jackknife and the Taylor linearization variance estimators are approximately the same for point estimates for which they are valid on large samples (i.e., they are asymptotically equivalent).

Hyunshik Lee

See also Balanced Repeated Replication (BRR); Finite Population Correction (fpc) Factor; Primary Sampling Unit (PSU); Taylor Series Linearization

Further Readings

Rizzo, L., & Judkins, D. (2004). Replicate variance estimation for the National Survey of Parents and Youth. *Proceedings of the Section on Survey Research Methods*

(pp. 4257–4263). Alexandria, VA: American Statistical Association.
Rust, K., & Rao, J. N. K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, *5*, 381–397.
Shao, J., & Tu, D. (1995). *The jackknife and bootstrap.* New York: Springer.
Westat. (2000). *WesVar 4.0, User's guide.* Rockville, MD: Author.
Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York: Springer.

Joint Program in Survey Methodology (JPSM)

The Joint Program in Survey Methodology (JPSM) is the oldest and largest program in the world offering graduate degrees in survey methodology. Located at the University of Maryland, JPSM was established in 1993 following the award of a grant from the National Science Foundation. JPSM was created to strengthen the federal statistical system by providing advanced training in survey statistics and methodology. It is a partnership between the federal government—specifically, the Interagency Council on Statistical Policy (ICSP)—and a consortium of the University of Maryland, the University of Michigan, and Westat. An innovative feature of the program is its sharing of classes across universities via video systems. In a typical term, almost half of JPSM's courses are shared this way, mostly with JPSM's sister program at the University of Michigan (the Michigan Program in Survey Methodology), but also with graduate programs at the University of Nebraska in Lincoln and the University of North Carolina at Chapel Hill. JPSM has become the hub of a national system of graduate education in survey methods.

Educational Programs at JPSM

JPSM offers a variety of educational programs including master's and Ph.D. programs in survey methodology, certificates in survey methodology and survey statistics, citations in introductory survey methodology and economic measurement, short courses, and summer internships. Its students include staff from the federal statistical agencies and from the many survey firms in the Washington, D.C., area that serve the statistical agencies. In addition, JPSM attracts and trains new entrants to the field of survey research.

JPSM began accepting students into the master's degree program in 1993. Because its primary goal is training survey practitioners, the master's program remains JPSM's central educational offering. As of September 2006, more than 140 students had graduated from JPSM with master's degrees. At that time, more than two thirds of the degree-seeking students were enrolled in the master's program. The master's program consists of a set of core courses, covering data collection, basic statistical methods, applied sampling, and total survey error. In addition, students get hands-on experience with surveys. Each year, the students conduct a survey, the JPSM Practicum; the master's students also offer survey consulting and are required to complete an internship at a survey firm. The master's program offers two "tracks," or concentrations—one in statistical science, the other in social science. Students in both tracks take the core courses. Those in the statistical science track take additional classes in probability and statistics, sampling, and survey estimation. Those in the social science track take additional courses in questionnaire design, cognitive processes related to surveys, and the analysis of survey data.

The purpose of the Ph.D. program is to train the next generation of researchers and teachers. The first doctoral students were accepted in 1999; as of September 2007, 13 students were pursuing Ph.D. degrees. Six doctorates have been awarded. Like the master's program, JPSM's Ph.D. program offers concentrations in statistical science and social science. Ph.D. students must meet three major requirements. After their first year, they are required to pass a qualifying exam. The qualifying exam covers the material in the required courses for the master's students. Doctoral students in statistical science are expected to be proficient in the topics covered in the core master's courses as well as the master's-level statistical science courses; similarly, doctoral students in the social science track must demonstrate their proficiency in the material covered in the classes required for the social science master's students. At the end of their second year, doctoral students must pass a comprehensive exam, demonstrating their mastery of even more advanced material. Finally, the doctoral students must complete a dissertation describing original research in survey statistics or survey methods.

JPSM also offers programs for students who are not seeking degrees. These programs are designed mainly to upgrade the skills of people who are already working as survey researchers. The most demanding are two certificate programs tailored to students who already have an advanced degree in another field but are seeking to learn more about survey methods. Each of the certificate programs is a bundle of six semester-length courses. In addition to its semester-length classes, JPSM also offers 20 or more short courses each year. The short courses are 1- or 2-day classes taught by experts on a given topic. Approximately 700 persons attend JPSM short courses each year. The citation programs recognize persons who have completed specific combinations of short courses and a semester-length class in survey methods.

JPSM also has one program—the Junior Fellows Program—designed for undergraduates; it seeks to recruit promising undergraduates to the field of survey research. The program consists of a weekly seminar at JPSM and an internship with one of the federal statistical agencies. In a typical year, about 30 students take part in the Junior Fellows Program; more than 200 Fellows have taken part since the program began.

Future Directions

As JPSM enters its second decade, it seeks to expand its ties to other programs (especially the Michigan Program in Survey Methodology), to become a more conventional department at the University of Maryland, and to continue to provide innovative educational offerings. JPSM's graduates have already made their mark on the federal statistical system and on the field more generally; their impact on survey research is likely to grow in the years to come.

Roger Tourangeau

Further Readings

Joint Program in Survey Methodology: http://www.jpsm.umd.edu

Journal of Official Statistics (JOS)

The *Journal of Official Statistics* (JOS) was launched in 1985. It is published by Statistics Sweden, the National Statistical Institute of Sweden. It replaced the then–century-old *Statistisk tidskrift* (Statistical Review). The ambition was to make JOS an internationally recognized communication medium on methodology and policy matters facing national statistical institutes and other producers of statistics. The language is English. The intended readers are working primarily at statistical agencies or in universities or private organizations and dealing with problems that concern aspects of official statistics and other production of statistics.

JOS functions as any other respected journal. All in-scope submissions are sent to referees for evaluation, and an associate editor together with the chief editor make the editorial decision for each manuscript. The number of associate editors has grown over the years from 5 in 1985 to 40 in 2007. The proportion of accepted manuscripts is approximately 22%. Examples of topics include new methods, interesting applications of known methods, comparative studies of different methods, authoritative expositions of methods in a certain field, and discussions of policy issues.

It is believed that the official statistics community has an unexploited potential for producing methodological articles. It seems that government statisticians publish less frequently than other professionals involved in statistics. Documents often remain at the draft stage without any international circulation. Sometimes these authors do not have to publish, and incentives to publish are not provided by their agencies. In consequence, many valuable contributions remain within the home-agency and do not initiate a more extensive research process involving scientists at other agencies, universities, or research organizations with an interest in these matters. JOS has been very successful as an outlet for these types of authors, and the journal's policy has been to provide them with as much guidance as possible, perhaps more than what is expected by scientific journals.

JOS is published four times a year. All issues from the start are accessible on JOS's Web site, which is searchable. All access is free of charge, all articles can be downloaded with no windows in place; that is, as soon as an issue has been printed it is available on the Internet. The Web site also contains information about subscription rates.

Lars Lyberg

Further Readings

Journal of Official Statistics: http://www.jos.nu

K

Key Informant

Within the context of survey research, *key informant* refers to the person with whom an interview about a particular organization, social program, problem, or interest group is conducted. In a sense, the key informant is a proxy for her or his associates at the organization or group. Key informant interviews are in-depth interviews of a select (nonrandom) group of experts who are most knowledgeable of the organization or issue. They often are used as part of program evaluations and needs assessments, though they can also be used to supplement survey findings, particularly for the interpretation of survey results. Key informants are chosen not because they are in any way representative of the general population that may be affected by whatever issue is being studied, but because they are believed to have the most knowledge of the subject matter.

Key informant interviews are especially beneficial as part of an initial assessment of an organization or community issue, allowing for a broad, informative overview of what the issues are. In survey studies, key informant interviews can be valuable in the questionnaire development process, so that all question areas and possible response options are understood. Relying on this method is also appropriate when the focus of study requires in-depth, qualitative information that cannot be collected from representative survey respondents or archival records. While the selection of key informants is not random, it is important that there be a mix of persons interviewed, reflecting all possible sides of the issue at study. Key informant interviews are most commonly conducted face-to-face and can include closed- and open-ended questions. They are often audio-taped and transcribed so that qualitative analyses of the interviews can be performed. Key informant interviews are rarely used as the sole method of data collection for a study or particular issue, as there is little generalizability that can come from them. However, they have a useful role, especially at the beginning stages of research studies when information gathering and hypothesis building are the goal.

Jennifer A. Parsons

See also Informant; Proxy Respondent

Further Readings

Johnson, J. (1990). *Selecting ethnographic informants.* Newbury Park, CA: Sage.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Werner, O., & Schoepfle, G. (1987). *Systematic fieldwork.* Newbury Park, CA: Sage.

Kish, Leslie (1910–2000)

Leslie Kish was a statistician, sociologist, and cofounder of the Institute for Social Research at the University of Michigan. His work had a profound and lasting effect on the field of survey research. His book *Survey Sampling*, published in 1965, formulated many of the principles that are today taken for granted in scientific survey research. The theory of equal probability sampling was first proposed in *Survey Sampling*, as was that of the design effect (*deff*). The Kish method of selecting respondents with equal probability is named for Leslie Kish. He was also an early proponent of counting and measuring nonresponse in survey research.

Kish emigrated to the United States from Hungary along with the rest of his family in 1925 at the age of 15. His father died shortly after the family arrived in the United States, and Leslie had to find work to help support the family. While working full-time he finished high school through an evening program in 1930. He then enrolled in the City College of New York while he continued to work full-time during the day. With less then 1 year left in his college studies, in 1937 he volunteered for the International Brigade and fought for the Spanish Loyalists in the Spanish Civil War.

In 1939 he returned to New York City and the City College, graduating with a bachelor's degree in mathematics later that year. He then took a job with the U.S. Census Bureau in Washington, D.C., and after working there a short time, he obtained a job with the U.S. Department of Agriculture (USDA). There he was the department's first sampling statistician. His boss at that time was Rensis Likert. Kish, Likert, and others at the USDA implemented the scientific survey sampling techniques that are still in use today. In 1942 Kish left the USDA to join the U.S. Army and served as a meteorologist until 1945.

In 1947 Kish and others (including Likert) went to the University of Michigan to establish the Survey Research Center, which went on to become the Institute for Social Research. At Michigan, Kish first earned a master's degree in mathematical statistics in 1948 and then a Ph.D. in sociology in 1952. He retired in 1981 and was professor emeritus at Michigan until his death in 2000. While in earlier jobs Kish was able to effect change at an organizational level, teaching at the University of Michigan provided him with a much larger stage from which he helped shape the entire field of survey research.

Kish was a devoted and enthusiastic scholar. He was a Fellow of both the American Academy of Arts and Sciences and the American Association for the Advancement of Science. In 1994 he was named Honorary Fellow of the International Statistical Institute, an honor that has been referred to as the Nobel Prize of Statistics. In 1995 Kish was named Honorary Member of the Hungarian Academy of Sciences.

He also had a passion for teaching and the dissemination of knowledge. Eager to see scientific sampling methodology spread to other countries, he started the Sampling Program for Foreign Statisticians in 1961. The program was an unqualified success. It has trained hundreds of statisticians from scores of countries. The program continues its mission today as the Sampling Program for Survey Statisticians.

Eric White

See also Design Effects (*deff*); EPSEM Sample; Equal Probability of Selection; Institute for Social Research (ISR); Kish Selection Method; Sample Design

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish Selection Method

Most surveys of attitudes and opinions have two stages: The first is drawing a random sample of dwelling units, and the second is selecting one person within a dwelling unit to interview. Leslie Kish published a landmark article in the *Journal of the American Statistical Association* in 1949 that described a rigorous, almost pure probability method of sampling persons within households to be surveyed. He named two basic conditions: (1) There must be a known probability of inclusion (excluding zero) of each adult in the population, and (2) it must be a practical and efficient procedure to implement.

Although survey data collection at that time often was conducted in person, Kish's plan is suitable for telephone surveys as well. It is accomplished by the interviewer's listing the age and sex of the adult members of the household and their relationship to the head of household and then consulting a table to choose the correct respondent. The interviewer numbers all the males first from oldest to youngest and then all the females in order of decreasing age. Kish provided two different sets of tables. In his first example, the interviewer has six tables, labeled A to F, which indicate first the numbers of adults in households from 1 through 6 and then below, the number of the adult to be interviewed. Each table lists the target individual in systematic order that differs among the tables, and the first three tables are set up in such a way that males are more likely to be selected because males tend to be underrepresented in surveys. In his second example, used more frequently in research today than the first procedure, there are eight tables labeled A, B1, B2, C, D, E1, E2, and F. Tables A, C, D, and F will be used one sixth of the time, and the others will be used one twelfth of the time. These tables give equal chances of selection to individuals in households of 1, 2, 3, 4, and 6 adults. Dwelling units with 5 adults are overrepresented to "compensate" for the inability of this method to represent households of more than 6 adults, a very small proportion of the population. Kish estimated, at the time he was writing, that only 1 or 2 adults per 1,000 would not be represented, usually young females.

Here is one example of the Kish method question wording, used by Robert W. Oldendick and his colleagues:

> *In order to select the right person to interview, we need to list all the people living in your household who are 18 years of age or older. First, could you tell me the ages of all the males living in your household who are 18 years of age or older—that is, from the oldest to the youngest? Next, could you tell me the ages of all the females living in your household who are 18 years of age or older—that is again, from the oldest to the youngest?*

Some researchers have introduced other methods because they believe that Kish's procedure takes too much time to enumerate household members, respondents may perceive it as intrusive, and it may increase refusal rates. In other words, they criticize it for potentially adding to nonresponse, although it can decrease within-unit noncoverage. The method may be more of a problem in larger, as opposed to smaller, households. Age is not asked in one-person households or in one male/one female units. Another criticism is that the tables may be outdated because of changes over time in the distribution of households or that the tables are inappropriate for some countries with different population patterns. Some surveyors have found that the Kish selection method, per se, is not very intrusive but that refusal rates tend to increase with any method that requires two consents—the first from the informant who answers the phone and the second from the designated respondent. Several researchers report that the Kish method is still a reliable and noninvasive method when interviewers are skilled and well trained in it. It is important that interviewers have faith in the method because they otherwise unconsciously may communicate negative attitudes toward a method, thereby decreasing the cooperation rate. Several comparative studies have included the Kish method with other methods on coverage, response, costs, and other characteristics.

Cecilie Gaziano

See also Last-Birthday Selection; Within-Unit Coverage; Within-Unit Selection

Further Readings

Binson, D., Canchola, J. A., & Catania, J. A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics*, *16*, 53–59.

Czaja, R., Blair, J., & Sebestik, J. P. (1982). Respondent selection in a telephone survey: A comparison of three techniques. *Journal of Marketing Research*, *19*, 381–385.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*, 380–387.

Németh, R. (2001). *Respondent selection within the household—a modification of the Kish grid.* Paper presented at the sixth Austrian, Hungarian, Italian, and Slovenian meeting of Young Statisticians, Ossiach, Carinthia, Austria. Retrieved August 19, 2006, from http://www.math.uni-klu.ac.at/stat/Tagungen/Ossiach/Nemeth.pdf

Oldendick, R. W., Bishop, G. F., Sorenson, S. B., & Tuchfarber, A. J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics*, *4*, 307–318.

O'Rourke, D., & Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, *20*, 428–432.

KNOWLEDGE GAP

This entry focuses on the concept of the knowledge gap, which is a specific hypothesis within the area of diffusion of knowledge, and subsequent derivations. A number of models of social change are based on the notion that change is a cumulative process. In such models, small changes result in differential rates of change for the social system—slow system change at first, followed by increasingly faster rates of change during the middle of the change process, followed by slowing rates of societal change at the end. This process is reflected in the familiar "S-curve" which shows an accumulative sum over time. Two prominent models that follow the logic and assumptions of the cumulative process perspective include models of *diffusion of technologies* and *diffusion of knowledge* in the social system. Originally proposed by Phil Tichenor and his colleagues in the early 1970s, the knowledge gap hypothesis predicts that as mass-mediated information enters a social system, certain segments of the population (such as those with higher socioeconomic status [SES]) acquire the information faster than the other population segments (those with lower SES). This process results in an increase rather than a decrease in the gap in knowledge between these two segments over time.

It is important to emphasize that the knowledge gap hypothesis is not about whether or not there is a gap between high and low SES segments of the population. Rather, the hypothesis concerns the widening of the gap over time. Actual tests of the knowledge gap hypothesis require data over time and are not prevalent in the scholarly literature. Studies of differences in knowledge levels for various segments of the population at one particular point in time might be better thought of as studies of knowledge differences or deficits (although they typically describe themselves as "knowledge gap" studies). The original knowledge gap study described the gap itself and took it as a given. The knowledge gap hypothesis is offered as an explanation as to how the gap might grow to the size it sometimes becomes when the same amount of information is available through media that penetrate most of society.

The original article by Tichenor and colleagues pointed out what researchers had known for years: that certain people expose themselves to certain kinds of information more, pay more attention, and retain more of it than do others. As a result, those who want to acquire more information will do so more quickly than those who do not. Some researchers have suggested that interest in the given topic under consideration is actually the important factor that determines the rate of information acquisition; others suggest that the important factor is motivation. Whatever the explanation for the phenomenon, empirical evidence is voluminous and can be found throughout social science research literature under such topics as "selective attention," "selective avoidance," "selective retention," and more general studies of learning from the media.

The primary contribution of the knowledge gap hypothesis is that, by describing the process as one that occurs over time, and by focusing on more macro concerns, attention was turned to the implications of this phenomenon for the social system rather than for individuals.

Knowledge gap research has been extensive and examined in a number of fields, including communications, political science, survey research, psychology, sociology, rural planning, forestry and wildlife management, health and medicine, biology, and numerous others. It is safe to say that hundreds of knowledge gap studies have been conducted in the intervening years, examining contingent or limiting conditions or aspects of the knowledge gap associated with particular media (the original article focused on newspapers), ways of limiting or countering the knowledge gap, or the implications of the knowledge gap for such behaviors as cancer prevention or political participation.

As was the case almost since the inception of the knowledge gap hypothesis, most research efforts on the hypothesis continue to be conducted at one point in time. These studies do not investigate the knowledge gap hypothesis, per se, but instead focus on one possible result of the hypothesis, a gap in knowledge levels between different segments of society at a particular time point. The most recent extensions of the knowledge gap have focused on the role of other media in supplementing knowledge or limiting the knowledge gap and have moved beyond newspapers to television and other media, spurred by the diffusion of new communication technologies such as computers and the Internet.

In addition to the examination of the knowledge gap hypothesis in relation to knowledge gain, the ideas developed in the knowledge gap hypothesis have been extended to access to traditional media and new technologies, resulting in several new research terms, such as the *digital divide*, *communication gaps*, and *communication inequities*. In studies of the digital divide, interest is often on diffusion and access to computers and the Internet and the implications of differential access for members of society.

Daniel G. McDonald

See also Level of Analysis; Longitudinal Studies; Panel Survey; Political Knowledge

Further Readings

Bonfadelli, H. (2002). The internet and knowledge gaps: A theoretical and empirical investigation. *European Journal of Communication*, *17*(1), 65–84.

Gaziano, C. (1983). The knowledge gap: An analytical review of media effects. *Communication Research*, *10*(4), 447–486.

Gaziano, C. (1997). Forecast 2000: Widening knowledge gaps. *Journalism and Mass Communication Quarterly*, *74*(2), 237–264.

Kwak, N. (1999). Revisiting the knowledge gap hypothesis: Education, motivation and media use. *Communication Research*, *26*(4), 385–413.

Liu, K., & Eveland, W. P. (2005). Education, need for cognition and campaign interest as moderators of news effects on political knowledge: An analysis of the knowledge gap. *Journalism and Mass Communication Quarterly*, *82*(4), 910–929.

McDevitt, M., & Chaffee, S. H. (2000). Closing gaps in political communication and knowledge. *Communication Research*, *27*(3), 259–292.

Tichenor, P. J., Donohue, G.A., & Olien, C. N. (1970). Mass media flow and differential growth in knowledge. *Public Opinion Quarterly*, *34*(2), 159–170.

Viswanath, K., Breen, N., Meissner, H., Moser, R. P., Hesse, B., Steele, W. R., et al. (2006). Cancer knowledge and disparities in the information age. *Journal of Health Communication*, *11*, 1–17.

Viswanath, K., & Finnegan, J. R., Jr. (1996). The knowledge-gap hypothesis: Twenty-five years later. *Communication Yearbook*, *19*, 187–227.

Knowledge Question

A knowledge question is designed to capture the extent to which people have stored factual information in long-term memory and how well they can retrieve and respond with that information when asked a survey question about a given topic. Knowledge, as a concept, is distinct from attitudes and opinions. Knowledge, moreover, is not synonymous with the term *information.* Whereas information describes a wider breadth of content that may include non-neutral as well as neutral elements, knowledge is considered neutral, factual content. The term *factual* is one that has been debated. Facts refer to those content elements on which there is consensual agreement. For example, it is a fact that it takes two thirds of Congress to override a presidential veto. Whether or not a given politician was arrogant during a debate, however, is a subject that is not likely to yield consensual agreement. Many topics can be assessed with knowledge questions, including politics, health, consumer issues, popular culture, science, and education.

Knowledge questions are used for several purposes. First, knowledge questions can be used to screen people for follow-up questions about people's attitudes on a given topic. If it is determined that a person has little or no knowledge of a topic, it may not be efficient for the researcher to ask that person follow-up attitude questions on that subject. Second, knowledge questions are used to assess a person's intellectual engagement on a topic, because knowledge is often predictive of various attitudes and behaviors. For example, citizens who are knowledgeable about politics tend to be the citizens who vote in elections; in other words, political knowledge predicts voting. Third, knowledge questions may be used for evaluative purposes. A school's performance may be assessed by how well its students have mastered a given set of material.

There are two primary types of knowledge questions: *recognition* and *recall.* Awareness questions are often assessed with recognition measures. Recognition measures ask people whether or not a given person, topic, or event is familiar to them. Specific knowledge content, however, is often assessed through recall measures. Recall items come in two forms: aided and unaided. An aided recall question is a closed-ended question that presents a respondent with several response choices from which the respondent selects an answer. An unaided recall question is asked in an open-ended format; the respondent receives no hints about the answer, and the respondent's answer is often recorded verbatim. Assessing a respondent's general awareness of a person, topic, or event is often less cognitively taxing on him or her than assessing a respondent's memory of specific knowledge content.

The main concern over recognition items is social desirability bias. People may say that they recognize an object because they want to appear informed. Consequently, some researchers follow up recognition items with specific content questions or include false (bogus) names or events in their recognition items to determine the extent to which people provide false positives (errors of commission) when it comes to recognition.

A general concern over asking knowledge questions is that respondents may feel intimidated when they perceive that they are being tested; they may be afraid of giving the incorrect response and looking ignorant or foolish. Anxiety can sometimes be minimized by prefacing questions with phrases such as *Do you happen to know* . . . or *As far as you know* . . . . Prefacing knowledge questions with softening phrases, however, may make people feel too comfortable giving "don't know" responses when they do know the correct answers but are not confident they know them and thus are hesitant to take a chance answering the questions. This can be problematic for researchers who are assessing the knowledge levels among different groups, particularly if some groups have greater propensities to guess than other groups do. For example, research has long shown that women are less likely to guess at knowledge items than are males. It is difficult to sort out a group's propensity to guess from its true levels of knowledge acquisition.

Knowledge levels for populations have been assessed with both direct and indirect measures. Direct measures ask specific, factual questions. Indirect measures include interviewer impressions of respondents' knowledge levels. While interview ratings have been used as knowledge measures, the validity of interviewer assessments is questionable because interviewer ratings are inherently subjective.

Mail and Internet surveys are problematic modes for conducting studies of people's knowledge levels, because respondents have the opportunity to look the answers up before completing the survey instruments. Consequently, telephone surveys and face-to-face interviews are preferable modes for assessing knowledge levels.

Kate Kenski

See also Aided Recall; Aided Recognition; Bogus Question; Closed-Ended Question; Errors of Commission; Open-Ended Question; Political Knowledge; Respondent Burden; Social Desirability; Unaided Recall; Verbatim Responses

Further Readings

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design—for market research, political polls, and social and health questionnaires.* San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

L

Language Barrier

Language barrier dispositions occur in U.S. surveys when a household member or the sampled respondent does not speak or read English (or another target language) well enough to complete the interview. The language barrier disposition is used in all surveys, regardless of the mode (telephone, in-person, mail, and Internet). Language barrier dispositions in surveys in the United States are not common, but their frequency is growing. Approximately 20% of the U.S. population in 2005 spoke a language other than English in their home, according to the U.S. Census Bureau. Furthermore, the 2005 U.S. Census estimates show upwards of 5 million residents being "linguistically isolated," in that they can speak little or no English. Language barriers are more likely to occur when data collection is conducted in central city areas and in rural areas of the Southwest.

The language barrier disposition functions as both a temporary and a final disposition. Cases may be coded temporarily with a language barrier disposition and then contacted again (in the case of a telephone or in-person survey) by an interviewer who speaks the same language as the household or the sampled respondent. In a mail or Internet survey, the survey organization may re-mail or resend a translated version of the questionnaire. However, cases with a language barrier final disposition often are considered eligible cases and thus are factored into survey nonresponse rates. The exception to this would be if the target population for a survey specified that respondents must speak English (or another specific language) to be eligible to complete the questionnaire.

Survey researchers may include a variety of categories within the language barrier disposition. One category of language barrier is used in telephone or in-person surveys for those cases in which there is no one in the household present at the time of contact who can speak or understand the language in which the introduction is spoken. Other categories of language barriers include cases in which the sampled respondent does not speak the language in which the interview is conducted or does not read the language in which the questionnaire is printed (for mail surveys) or displayed (for Internet surveys). Finally, a third category of language barrier occurs in in-person and telephone surveys when an interviewer fluent in the language spoken by the household is not available to be assigned to the case at the time of contact.

Because cases with a language barrier disposition increase the nonresponse rates in a survey, researchers fielding surveys in areas known to be multi-lingual often use practical strategies to ensure that households or respondents who do not speak the language in which the interview is to be conducted can complete the survey. In telephone and in-person surveys, these strategies usually include employing multi-lingual interviewers and arranging to have the survey questionnaire and all supporting materials translated into one or more additional languages. In mail and Internet surveys, these strategies usually include having the questionnaire and supporting materials translated into one or more additional languages and then re-sending (via U.S. Postal Service or email, depending on the mode) the translated questionnaire and materials to respondents who reported that they are unable to read English.

Finally, it is worth noting that occasionally a language barrier is used by a household or a respondent as an excuse to avoid completing the interview. These cases are very rare and happen mostly in in-person and telephone surveys. If this situation occurs, it should be considered a refusal, not a language barrier. Because there is no standardized manner for interviewers to determine if this situation has occurred, these types of cases should be reviewed carefully by a supervisor or contacted again by a different interviewer.

Matthew Courser

See also Dispositions; Final Dispositions; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

LANGUAGE TRANSLATIONS

Many survey projects at national and cross-national levels use questionnaires developed in one language and translated into another. The quality of these translations is a crucial factor in determining the comparability of the data collected. Conversely, poor translations of survey instruments have been identified as frequent and potentially major sources of survey measurement error. This entry outlines key aspects of conducting and monitoring survey language translations.

Essentially, translation allows researchers to collect data from people who cannot be interviewed in the language(s) in which a questionnaire is already available. In countries such as the United States, long-standing linguistic minorities and newly immigrated groups make translation into multiple languages essential to ensure adequate coverage and representation of different segments of the national population. The 2000 U.S. Census was available in six languages, with language aids being provided for 49 languages. The 2010 Census is likely to accommodate even more languages. In Switzerland, different segments of the population only speak one of the three main official languages. It is thus necessary to field Swiss national surveys in Italian, French, and Swiss German. At the same time, translation is occasionally motivated by official rather than practical requirements; in Great Britain, for example, some surveys are translated into Welsh, although most of the Welsh population uses English with native language competence.

Translation is not always part of an original survey research plan; sometimes it has to be added at short notice in order to interview unanticipated linguistic groups. Questionnaires or sets of questions are also sometimes simply "borrowed" from one study for use in another. In being borrowed for a new population, they may also need to be translated. Medical, psychological, and educational diagnostic instruments, often developed at great cost, are regularly adapted for new locations on the basis of translation. The aim may simply be to use the questions, not to compare data across populations. These are usually proprietary instruments for which permission is required and a fee possibly levied. In the cross-national context, international surveys may occasionally be able to use English as a lingua franca and thus avoid translating. However, most multi-lingual and cross-national projects must rely on translation to produce questionnaires in the languages needed.

Study Design and Translation

Two basic models are used to conduct multi-lingual studies. Researchers can decide to ask the same questions or to ask questions that differ in content but are thought to be comparable across populations. Ask-the-same-question models are by far the more common. In these models, a questionnaire is developed in one language, which is then translated to produce the other language versions required. Following terminology used in the translation sciences, the language translated "out of" is known as the *source language* and the language translated "into" is known as the *target language*. Most multi-lingual surveys, whether national or international, adopt ask-the-same-question procedures. If developed and implemented properly, they permit researchers to compare data, variable by variable, across populations.

Managing Translation Efforts

How a study conducts its translation effort depends on budgets, time schedules, the languages and number of respondents involved, and locations to be covered. Equally important are the research team's views on best practice and thus on whether participants are required to adhere to specifications or guidelines. Some projects delegate responsibility for translations to their fielding organization(s) or to translation bureaus or brokers. These may take over the translation effort or collaborate with the project staff or researchers, providing translators and possibly translation software, but leaving quality assessment up to the project staff. Alternatively, project staff can hire or buy translation tool and memory software and manage their own efforts, working with in-house and contracted personnel as needed. Some organizations have in-house language units or regular contractors for target languages they frequently need. In oral translation, translation call centers may be involved.

Technical Tools for Production and Assessment

Driven by time constraints, cost-efficient production and monitoring methods, as well as project volume and complexity, technology plays an increasingly important role in survey translation. Various technical aids are available, including translation tools, vocabulary databases, and project management software to monitor delivery schedules, phase completion, and documentation. Research is under way in several organizations to develop tools tailored to the needs of survey translation and assessment in the context of computer-assisted applications.

Translation Monitoring and Documentation

Systematic monitoring of translation quality also calls for technical support. In addition, team translation efforts rely on documentation from one stage for efficient execution of the next. Numerous projects use templates aligning source and target texts to facilitate the translation and documentation processes. In some projects, final draft versions are monitored centrally; in others, this responsibility is left with the individual language teams. Those translation teams who are monitoring these processes centrally need to align and compare the matching versions of different languages and any accompanying documentation. When source questionnaires are altered over time, monitoring also involves identifying and aligning different versions of the source and translated questionnaires and any notes on harmonization made across shared languages.

No translation or monitoring procedure is an automatic guarantee of quality. Even when guidelines or protocols are appropriate, errors will occur. An optimal model of quality assurance for survey research is still being developed. Careful implementation and revision conducted by appropriate personnel is critical. After the translation team considers their work completed, an external quality assessment is recommended. Models for external assessments are currently being developed in the international context.

Harmonization

As used here, the term *harmonization* is an extension of the principle of asking the same questions. In multilingual studies, several countries or groups may use or share the same language (e.g., Spanish or French). Sometimes a study stipulates that efforts must then be undertaken to harmonize questions across those populations sharing a language. In a harmonizing procedure, countries sharing a language compare and discuss their individually translated versions with a view to removing any unnecessary differences in a final harmonized version. The degree to which harmonization is obligatory differs; many studies recognize that regional standards of a shared language result in necessary differences in translations.

Team Translation

Views on what counts as good survey translation practice have changed noticeably in the last decade. Translation guidelines produced by the European Social Survey, by the U.S. Bureau of Census, by the World Health Organization's World Mental Health Initiative, and by the International Workshop on Comparative Survey Design and Implementation, all available online, emphasize the benefits to be gained from organizing survey translations in a team effort.

These team procedures consist of (at least) five steps that may be reiterated: (1) draft *translations* by translators; (2) *review* by the entire team; (3) *adjudication* by the team or a subgroup of it; (4) *pretesting* and translation adjustment; and (5) *documentation*.

In these procedures, acronymed in the European Social Survey guidelines as TRAPD, documentation is an ongoing process, informing each stage of development and documenting final outcomes and decisions.

Team translation efforts bring together a group of people with the skills to resolve the challenges inherent in producing good survey translations. In such arrangements, several translators produce draft translations. These translators are people with the relevant language competence, training in translation, and recent practice in translation work. Their translations are revised and finalized in collaboration with substantive experts and survey research specialists. Documentation of, and commentary on, output facilitates each stage. Team approaches emphasize the need for a collaborative effort of people with different skills. Because they incorporate review into the translation process, translations are reviewed and corrected before the final version is approved.

Back Translation

A more traditional form of survey translation usually has one translator translate. Assessment of the translation is often made by having a second translator produce a *back translation* of the translated text. Thus, if an English questionnaire is translated into Czech, the back translation step translates the Czech translation into English, without the translator seeing the original English version. The assessment of the Czech text is made by comparing the two English versions (source and back translation). If these are thought to be comparable (enough), the Czech translation is considered to have passed muster.

Back translation was one of the earliest procedures to establish itself in survey research and, over time, became associated with quality assurance. In translation studies, however, translation quality is normally assessed by focusing on the target translation, not source-language texts. Back translation can be an aid to researchers who do not understand the target language and want to gain a sense of what is in a text. However, as increasingly recognized, back translation cannot function as a refined tool of quality assessment.

Translation: Difference and Sameness

All translation is expected to convey in the target language whatever meaning is considered essential in the source text. However, the purpose of a translation determines what counts as an essential, and the target audience and the mode of presentation further affect how essentials are realized in translation. The extensive debate in translation theory literature related to these issues is beyond the scope of the present entry.

Behind any decision to translate survey questions lies the assumption that the translation will result in the same question in a different language. This is not a trivial assumption. It implies that we can identify those constituents of a question that must be translated to assure that the same question results. Currently we do not have a fully fledged and robust theory or even a description of what such constituents might be. In practical terms, most survey translations concentrate on rendering the semantic content of the source questions. Thus, if a source question asks *How often do you visit your parents?* a translation would be expected to refer to the set of people normally associated with the term *parents* in English and not, for example, to *relatives* and to ask a question about *visits* rather than a question about, for example, *seeing*.

Cultural and linguistic considerations often interact to complicate translation. A literal translation of the question *How often do you visit your parents?* would puzzle populations that commonly share living arrangements with their parents. It might also be necessary to specify linguistically which *parents* are intended and who *your* should refer to.

Adaptations

Translation alone is often not sufficient to render questions appropriately and adaptations may be required. *Adaptations*, as the term is used here, are changes that are not purely driven by language. They take different forms, partly determined by the needs or context of a given target group. Some adaptations are simple, such as changing Fahrenheit measurements to Centigrade. Others are more complex, depending also on the purpose of the question. Questions on health, for instance, may need to mention different symptoms for various cultures to determine the presence of the same complaint or illness across populations. In tests of knowledge, different questions may be needed to avoid favoring one population, and in skills and competence research, target-language versions depend heavily on adaptation. Visual components may be adapted to accommodate the direction in which material is read in a given culture or language. Some populations may be unfamiliar with questionnaires and need more navigational guidance and instructions. Careful adaptation helps researchers produce target-culture questions that will collect data comparable to that collected with the source questions.

Oral Translation and Interpreting

Sometimes surveys do not produce written translations but ask bilingual interviewers to translate orally from source-language questionnaires while conducting the interview. The limited research available on oral translation points to multiple differences between orally translated and written translations, thus raising questions about quality and comparability and, in turn, the reliability of the resulting data. Interpreted interviews, in which an interpreter mediates between interviewer and respondents, are a second underresearched form of oral translation. However, if a language does not have a written form, oral translation and interpreting are the only ways to interview respondents who understand only that language. Guidelines on interpreting are currently being developed by the U.S. Bureau of Census and the International Workshop on Comparative Survey Design and Implementation.

Janet Harkness

See also Bilingual Interviewing; Language Barrier

Further Readings

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: Wiley.

Harkness, J. A. (2007). Improving the comparability of translations. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79–94). London: Sage.

Harkness, J. A. (2008). Comparative survey research: Goals and challenges. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 56–77). New York: Psychology Press.

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Questionnaire translation and assessment. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Hoboken, NJ: Wiley.

Harkness, J. A., Schoebi, N., Joye, D., Mohler, P., Faass, T., & Behr, D. (2008). Oral translation in telephone surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 231–249). Hoboken, NJ: Wiley.

Pan, Y., & de la Puente, M. (2005, August 24). *Census Bureau guidelines for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed*. Washington, DC: U.S. Bureau of the Census, Statistical Research Division. Retrieved January 12, 2007, from http://www.census.gov/srd/papers/pdf/rsm2005-06.pdf

Last-Birthday Selection

Survey researchers are usually concerned with choosing respondents within households after households are selected randomly. Within-unit coverage and nonresponse are key issues, so researchers want to select the correct respondent and gain his or her cooperation. Each of these goals has costs. One popular quasi-random compromise is the last-birthday (LB) method of selecting respondents from within a sampled household in random-digit dialing surveys. It circumvents the pitfalls of pure or nearly pure random methods by being relatively quick, easy, and likely to secure cooperation. Probability methods can involve a potentially lengthy and intrusive process of querying the informant (person who answers the phone) about all household members eligible to be interviewed before selecting the correct respondent from the resulting list.

An example of LB question wording is, *In order to determine whom to interview, could you tell me, of the people who currently live in your household who are 18 or older—including yourself—who had the most recent birthday? I don't mean who is the youngest adult, but rather, who had the most recent birthday?* If the respondent does not know all the birthdays, the following question can be asked: *Of the ones you do know, who had the most recent birthday?* Some researchers have sought to study the best wording to choose the correct respondent and secure cooperation.

The first published description of birthday methods tested a next-birthday (NB) method, assuming that the incidence of births is random and the first stage of selection in a two-step process. The researchers, Charles Salmon and John Nichols, considered this method to be the second stage of sampling, with all members of a household having an equal probability of being chosen. After implementation, however, the researchers recommended use of a last-birthday (or most recent birthday) procedure because some respondents were confused about whom to nominate. LB now is used more frequently than NB because it appears that informants understand it better or are more likely to know past birthdays than upcoming birthdays.

Several studies show that births are not necessarily distributed randomly across months in some countries and that LB also tends to produce distributions of birthdays skewed toward the months closely preceding the interviews in households of two or more adults. Skewed distributions should not result in bias, however, unless birthdays are connected to interview topics. Another drawback is that informants in large households, compared with those in small households, are less likely to know everyone's birthday. LB tends to overrepresent females but may represent blacks and younger males better than do other procedures. Females may be overrepresented because they tend to answer the phone more frequently than males, and sometimes they may be "protecting" males from coming to the phone. Other times informants want to do the interview so they inaccurately report they had the last birthday.

It is important that informants understand the question because misunderstanding can lead to within-unit coverage problems, especially among the less educated, the foreign born, and larger households. Sometimes informants erroneously think that the interviewer wants the youngest household member. Interviewers also should ascertain the respondent's day and month of birthday to be sure they are speaking with the correct person. Stressing the need for scientific accuracy can improve the response rate. Research on the LB method indicates that the respondent has been nominated incorrectly between about 10% and 25% of the time. Intensive training of interviewers regarding the question and accuracy of respondent selection, especially in households of three or more persons, is likely to decrease the incidence of coverage mistakes. Some research on the LB technique shows that the inaccurately selected respondents did not contribute to biased results, whereas other research demonstrates some contribution to bias when certain demographic variables associated with incorrect respondent selection were related to specific substantive results.

LB can be more economical than, or comparable to, the Kish respondent selection technique (which is close to being a pure probability method) in length of contact with informants, interview length, and cooperation rates. LB can be a little less economical than nonprobability methods, including those that simply select the most available or willing respondent. This is because any technique that selects one and only one respondent in a household requires many more callbacks, on average, to contact that respondent than a technique that allows essentially anyone to be the respondent.

Cecilie Gaziano

See also Informant; Kish Selection Method; Within-Unit Coverage; Within-Unit Selection

Further Readings

Binson, D., Canchola, J. A., & Catania, J. A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics*, *16*, 53–59.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Lavrakas, P. J., Bauman, S. L., & Merkle, D. M. (1993). The last-birthday selection method and within-unit coverage problems. *Proceedings of the Section on Survey Research Methods* (pp. 1107–1112). Alexandria, VA: American Statistical Association.

Lavrakas, P. J., Stasny, E. A., & Harpuder, B. (2000). A further investigation of the last-birthday respondent selection method and within-unit coverage error. *Proceedings of the Section on Survey Research Methods* (pp. 890–895). Alexandria, VA: American Statistical Association.

Oldendick, R. W., Bishop, G. F., Sorenson, S. B., & Tuchfarber, A. J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics*, *4*, 307–318.

O'Rourke, D., & Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, *20*, 428–432.

Salmon, C. T., & Nichols, J. S. (1983). The next-birthday method of respondent selection. *Public Opinion Quarterly*, *47*, 270–276.

LEANING VOTERS

Leaning voters is a term in politics and survey research methods that has several meanings. The nominative application refers to voters who are not strongly affiliated with any political party, nor are they hard-core independents. They lean toward being a partisan of one stripe or another.

Another definition refers to voters who do not indicate that they are solidly supporting one candidate or another—but they do lean toward supporting a candidate. Related to this, a third definition is a verb: Researchers *lean* voters when they allocate undecided voters to one candidate or another through use of various heuristic or empirical methods. The term derives from how the two concepts are traditionally measured.

Candidate Support Measures

When researchers measure candidate support in pre-election polls, there usually are respondents who initially support a candidate and those who say they are undecided (or refuse to tell the interviewer). Many researchers follow up with these "no opinion" respondents, probing to determine whether they lean toward supporting a candidate. Then these leaners are combined with the firmer supporters, that is, the choosers, to report total support. Experience has shown that this is the most accurate way of determining candidate support. Using this measurement method, researchers can separately analyze the truly undecided, a candidate's strong supporters, and those who are leaners.

Party Identification Measures

Because party affiliation is a psychographic attitude or orientation, rather than a hard-and-fast demographic characteristic such as gender, age, race, or educational attainment, different partisanship measures can have a real effect on the proportion of partisans in polls and other sample surveys. There are several standard measures that enable researchers to partition U.S. voters into a continuum ranging from "strong Democrat" through "independent" to "strong Republican." One standard measure of party identification was developed by researchers at the University of Michigan for the National Election Studies:

> *Generally speaking, do you consider yourself a Republican, a Democrat, an independent, or what?*
>
> If respondents choose a party, they are asked if they consider themselves *a strong [Democrat/Republican] or a weak [Democrat/Republican].*
>
> Respondents who say they are independents are asked, *Do you think of yourself as closer to the Republican or Democratic Party?*

It is these independents—the ones who choose Republicans or Democrats at this point—who are typically labeled "leaners."

The Gallup Poll uses a similar measure, but frames the initial question, *In politics today* . . . . Most researchers now agree that the Gallup measure allows for more short-term variability in party identification. Other polls use variations of these two methods. For example, the Minnesota Poll asks everyone who does not initially choose a party whether they lean toward the Democratic Party or toward the Republican Party; this is simpler and quicker and has the effect of providing a somewhat less cumbersome 5-point scale and does not partition partisans into weak and strong.

However, limiting party affiliation measures to closed-ended questions also may shunt those who consider themselves aligned with third parties into the leaner or independent categories. In some cases, especially in states where there are strong third parties (e.g., Minnesota), or in national elections when there is a strong independent party candidate (e.g., Ross Perot in 1992), these closed-ended questions may not allow researchers the analytical power they need to understand the electorate properly.

This measurement issue is key to understanding the research about the behavior, demographics, and attitudes of leaning voters compared with true partisans and independents. Earlier in the 20th century, some scholars argued that leaners were similar to independents, and a good way to analyze U.S. political data was to talk about a trichotomy—Democrats, Republicans, and everyone else. This ignored third-party candidates, such as George Wallace in the 1968 presidential election, Ross Perot in the 1992 presidential election, and others in various state elections. More recently, some scholars have made a strong empirical argument that voters who lean toward one party or another ultimately behave more like true partisans than independents, who are less likely to participate in politics and more likely to have fewer strong attitudes about public policy issues. They also argue that true independents are a small—about 1 in 10—portion of the electorate. Practically, however, many, if not most, media pollsters default to the earlier typology when they report and analyze poll results: Democrats, Republicans, and everyone else.

In election forecasting that uses pre-election polls, especially in close elections, what pollsters do with leaning voters when making their forecasts often will contribute to a correct or incorrect prediction. As such, understanding how leaning voters are measured is key to journalists, media readers and viewers, and other political analysts.

Robert P. Daves

See also Closed-Ended Question; Election Polls; Gallup Poll; Horse Race Journalism; National Election Studies (NES); Psychographic Measure; Undecided Voters

Further Readings

Green, D. P., & Schickler, E. (1993). Multiple-measure assessment of party identification. *Public Opinion Quarterly*, *57*(4), 503–535.

Keith, B. E., Magleby, D. E., Nelson, C. J., Orr, E., Westlye, M. C., & Wolfinger, R. E. (1992). *The myth of the independent voter*. Berkeley: University of California Press.

Mystery Pollster: http://www.mysterypollster.com/main/2005/03/realignment_or_.html

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Level of Analysis

A social science study using survey data can be set at the micro level when individuals are analyzed, or it can be set at a higher, more macro level when aggregates of individuals such as households, wards, precincts, firms, neighborhoods, communities, counties, provinces, states, or nations become the unit of analysis. This structural level, spanning the range from most micro to the most macro, at which a social scientific investigation is carried out is called *level of analysis*. A particular study may also cut across several levels of aggregation. For example, a multi-level study of the educational effectiveness of a certain education program may include pupil-specific, classroom-specific, school-specific, and school-district-specific information and analyze the data at each and all of the levels.

The choice of level of analysis should be driven by researchers' theory and, subsequently, their research questions. There are two large, contrasting issues of concern over why the level of an analysis must be carefully chosen and specified. The first is the famous issue, or the infamous problem, of the *ecological fallacy*, popularized by William S. Robinson in 1950. Simply stated, the ecological fallacy is an incorrect inference about individual or micro-level effects or relationships drawn by analyzing aggregate or macro-level data. Many theories are set at the individual level. However, it is easy to overlook the possible fallacy and study social relations in the aggregate because data are more widely available at that level.

The second issue is that of *emergent property*, which may appear when a number of simple entities (or individual actors or agents) operate in an environment, social or otherwise, forming more complex behaviors as a collective. Emergent properties are not reducible to the properties of the individual agents. This idea is attributed to Émile Durkheim in *The Rules of the Sociological Method*, initially published in French in 1895. The idea of emergent property is a potent and power one, and its influence can be found outside of the social sciences today. For example, researchers of artificial intelligence study the so-called emergent functionality. Put another way, a component has a particular functionality, which is not recognizable as a subfunction of the global functionality. For survey researchers, data collected at the individual level should not be aggregate in order to draw inference for a particular behavior at a higher level, which may be emergent.

Both the ecological fallacy and emergent property are important issues for survey researchers because the (primary) sampling unit of a survey sets a limit for the level of analysis a researcher wants to use. A sampling unit is the elementary unit that is sampled or selected for detailed examination, and valid statistical sampling requires that each sampling unit have a determinable nonzero chance of selection and that each be selected randomly. Statistical properties aside, sampling unit gives the level at which detailed information is acquired. For example, the General Social Survey (GSS) in the United States samples English-speaking individuals 18 years or older living in noninstitutional arrangements in the United States. Naturally, the GSS is most appropriate for analysis at the individual level.

All surveys on which individuals are interviewed are not at the individual level. For example, the Panel Survey of Income Dynamics (PSID) has household as the sampling unit. Even though it is the individuals who are interviewed, detailed information is available at both the individual and the household level. Similar surveys include the British Household Panel Survey (BHPS) and the European Community and Household Panel (ECHP). An analysis of the BHPS, ECHP, or PSID households should be as straightforward as an analysis of individuals using the GSS. An analysis of individuals from the BHPS, ECHP, or PSID, however, becomes trickier than one may think even though there is detailed information available for these individuals who are drawn from the households sampled by certain statistical principles because these individuals in the same household are no longer independent observations; they form clusters that deserve special statistical treatment.

Another example of surveys where the level is not set at the individual is the National Organizations Study (NOS). Even though the eligible establishments (or organizations) were initially identified through the 1991 GSS, the NOS sampled work organizations, with establishment as the sampling unit. Obviously, researchers analyzing the NOS data set the level of analysis at the organizational level. Clearly, the questions that get asked on a survey like the NOS will not be the same as those directed toward individual solo entrepreneurs because of emergent property in work organizations.

Finally, it is worth reiterating that the level of one's theory and that of the analysis must be consistent with each other. The proliferation of household-level surveys, together with the frequency in which data from such surveys are analyzed to answer individual-level questions, poses an interesting challenge for the survey researcher: Can household surveys be used to answer individual questions? The answer is a qualified "yes." The qualification comes in two ways: To the extent that dependence among members of the same household is recognized and handled properly, and to the extent that individuals' representation of the population is appropriate (if necessary, weighted), household surveys can answer individual questions because individuals are the ones who are interviewed individually, not collectively as a household. The last point also raises another question for survey researchers: How well can individuals represent households of which they are a member? That is a question for survey designers, and as such, it is beyond the issue of level of analysis.

Tim F. Liao

See also Ecological Fallacy

Further Readings

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). Boston: McGraw-Hill.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

Smith, P. B. (2002). Levels of analysis in cross-cultural psychology. In W. J. Lonner, D. L. Dinnel, S. A. Hayes, & D. N. Sattler (Eds.), *Online readings in psychology and culture* (Unit 2, chap. 7). Bellingham: Western Washington University, Center for Cross-Cultural Research. Retrieved March 12, 2008, from http://www.ac.wwu.edu/~culture/smith.htm

Level of Measurement

Level of measurement refers to the relationship between the numeric values of a variable and the characteristics that those numbers represent. There are five major levels of measurement: nominal, binary, ordinal, interval, and ratio. The five levels of measurement form a continuum, because as one moves from the nominal level to the ratio level, the numeric values of the variable take on an increasing number of useful mathematical properties.

Nominal

Nominal variables are variables for which there is no relationship between the numeric values of the variable and characteristics those numbers represent. For example, one might have a variable "region," which takes on the numeric values 1, 2, 3, and 4, where 1 represents "North," 2 represents "South," 3 represents "East," and 4 represents "West." Region is a nominal variable because there is no mathematical relationship between the number 1 and the region North, or the number 2 and the region South, and so forth.

For nominal variables, researchers cannot compute statistics like the mean, variance, or median because they will have no intuitive meaning; the mode of the distribution can be computed, however. Nominal variables also cannot be used in associational analyses like covariance or correlation and cannot be used in regressions. To use nominal variables in associational analyses, the nominal variable must be separated into a series of binary variables. Only nonparametric statistical tests can be used with nominal variables.

Binary

Binary or "dummy" variables are a special type of nominal variable that can take on exactly two mutually exclusive values. For instance, one might have a variable that indicates whether or not someone is registered to vote, which would take on the value 1 if the person is registered and 0 if the person is not registered. The values are mutually exclusive because someone cannot be both registered and not registered, and there are no other possibilities. Like with nominal variables, there is no mathematical relationship between the number 1 and being registered to vote, but unlike nominal variables, binary variables can be used in associational analyses. Technically, only nonparametric statistical tests should be used with nominal variables, but the social science literature is filled with examples where researchers have used parametric tests.

Ordinal

Ordinal variables are variables for which the values of the variable can be rank ordered. For instance, a researcher might ask someone their opinion about how the president is doing his job, where *1 = strongly approve, 2 = somewhat approve, 3 = somewhat disapprove*, and *4 = strongly disapprove*. In this case, the values for job approval can be ranked, and researchers can make comparisons between values, for example, saying that someone who gives a job approval value of 1 approves of the president more than someone who gives a job approval value of 3.

However, a researcher cannot make exact mathematical comparisons between values of the variable; for example, it cannot be assumed that a respondent who gives a job approval of 4 disapproves of the president twice as much as someone else who gives a job approval of 2. Researchers can, however, compare values using "greater than" or "less than" terminology and logic.

The mode and the median can be computed for an ordinal variable. The mean of an ordinal variable is less meaningful, because there is no exact numerical "distance" between the number assigned to each value and the value itself.

Ordinal variables can be used in associational analyses, but the conclusions drawn are dependent upon the way that numbers were assigned to the values of the variable. For instance, reassigning the values of job approval such that "strong approval" is now a 5, "somewhat approval" becomes a 4, and so on, would change the sign of the correlation between job approval and another variable. Thus, the associational relationship observed between two variables is a by-product of both the way the ordinal variables were coded and the underlying relationships in the data. Technically, only nonparametric statistics should be used with ordinal variables, but the social science literature is filled with examples where researchers also have used parametric statistics.

Interval

With interval variables, distances between the values of the variable are equal and mathematically meaningful, but the assignment of the value zero is arbitrary. Unlike with ordinal variables, the differences between values assigned to the variable are meaningful, and researchers use the full range of parametric statistics to analyze such variables.

As with ordinal variables, interval variables can be used in associational analyses, but the conclusions drawn are dependent upon the way that numbers were assigned to the values of the variable. Interval variables can be rescaled to have a different value arbitrarily set to zero, and this would change both the sign and numerical outcome of any associational analyses. Parametric statistics can be used with interval variables.

Ratio

With ratio variables, distances between values of the variable are mathematically meaningful, and zero is a nonarbitrarily assigned value. Anything that can be counted—votes, money, age, hours per day asleep—is a ratio variable.

Values assigned to ratio variables can be added, subtracted, multiplied, or divided. For instance, one can say that a respondent who views 6 hours of television per day views twice as many hours as another respondent who views only 3 hours, because for this variable, zero is nonarbitrary. By contrast, one cannot say that 60 degrees feels twice as warm as 30 degrees, because 0 degrees is an arbitrary construct of the temperature scale.

With ratio variables, researchers can calculate mean, median, mode, and variance and can use ratio variables in the full range of parametric associational analyses, with meaningful results.

Amy R. Gershkoff

See also Attitude Measurement; Interval Measure; Mean; Median; Mode; Mutually Exclusive; Nominal Measure; Ordinal Measure; Ratio Measure; Variance

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.
Fowler, F. J. (2002). *Survey research methods*. Thousand Oaks, CA: Sage.

LEVERAGE-SALIENCY THEORY

Leverage-saliency theory, as first proposed by Robert M. Groves and his colleagues in 2000, is a unifying theory to help explain survey nonresponse, with the goal of helping to identify strategies to counter nonresponse.

Nonresponse is a critical challenge to survey research. Those who do not respond to surveys (or to parts of questionnaires) may differ in important ways from those who do respond. Leverage-saliency theory attempts to describe the underpinnings of individual behavior related to the individual's choosing to cooperate or not to cooperate with a survey request. The theory posits that different people place a different level of importance to various attributes associated with a survey request. These attributes are like weights on a scale, tipping the scale to the sample person either acceding to or declining a particular survey request. An implication of the theory is that the response propensity of any one person deciding to cooperate or not with a specific survey request will vary across different survey requests and that few people will always agree or never agree to participate when they are sampled for a survey.

This entry describes several key attributes of the survey request and how these interact in terms of their leverage, value disposition, and saliency to affect survey response. The entry includes suggestions as to how interviewers can tailor their requests to make more salient those attributes with the greatest amount of positive leverage for an individual from the sample. This entry describes the leverage-saliency theory in terms of an individual sample person. In theory, the ability to alter the survey request design in ways that make salient those request attributes, to which various subgroups give positive leverage, will aid in increasing response among subgroup sample members as well. However, in practice, this often is hard to achieve at a cost-acceptable level.

Leverage-Saliency: Tipping the Scale

It can be helpful to think of survey request attributes as weights on a scale. Each attribute has three qualities: (1) the distance to the scale's fulcrum (point of balance), (2) the sample person's disposition toward the attribute (ranging from positive to negative), and (3) the saliency of the attribute. The further the attribute of the survey request is to the scale's fulcrum, the greater the amount of leverage it exerts in the sample person's decision making. A strong amount of leverage for a survey attribute (e.g., the perceived value of an incentive) is helpful only if the disposition toward this attribute is positive. If the sample person perceives the attribute as negative (e.g., being insulted by being offered too low an incentive), then the increased leverage of the attribute may decrease the likelihood of responding. The leveraged force from any single request attribute can be exerted only when that attribute is made salient in the sample person's decision-making process. Thus, the goal of the survey researcher is to make salient those attributes that have the greatest amount of positive leverage for a sample person. This also holds for any subgroup of the sample (e.g., 18- to 34-year-olds) for whom the survey researcher makes salient some request attribute believed to be positively valued by the members of that subgroup.

In theory, researchers who are able to recognize and make salient the survey request attributes that a sample person values positively increase the likelihood of the sample person cooperating with the survey request, thereby increasing the response rate. The challenge is that the level of importance (i.e., the leverage) of the different request attributes that the sample person views as positive, in most cases, is unknown to the researcher. In addition, sample members may hold several different survey request attributes with varying leverages and dispositions toward these attributes in the balance of their decision-making process. Taken together, this causes survey researchers to face a serious dilemma in trying to stem nonresponse.

Common Attributes of Survey Requests

In almost all survey requests, the researcher presents several attributes to the sample person. First, the survey researcher typically makes the topic of the survey known (i.e., makes the topic salient) to the sample person early in the survey request. If the topic of the survey holds low leverage for the sample person, this may not be enough to tip the scales in favor of the person's acceding to the request. However, other attributes of the request may contribute to the decision-making process of the sample person. These additional attributes include the nature and amount of incentive for responding and other persuasive communications, for example, being told that the survey will help one's local government decide about how to improve recreational opportunities in one's community. By making one or more of these other attributes salient, the survey researcher's efforts may produce a cooperating sample person. The saliency of these other request attributes, however, must outweigh the sample person's perceived burden of completing the survey.

Another survey request attribute is the authority and prestige of the sponsoring organization. The perception that the sample person has of the sponsoring organization can vary widely. Some subgroups of respondents may view a survey request from the government or from an academic institution as legitimate, and the survey's saliency may contribute to tilting the scale toward their responding. On the other hand, respondents who are members of subgroups that have been alienated by the sponsoring organization may perceive the authority of the request with great suspicion or hostility. In this case, researchers who emphasize the authority of the sponsoring organization inadvertently may be making salient an attribute that may tip the scale toward not responding.

Survey researchers also use incentives in survey requests to encourage a sample person to respond. If a sample person views an incentive positively and with a great amount of leverage, making an incentive salient in the survey request may be the difference between a refusal and an acceptance to respond. Past research has found that the effects of incentives on survey response vary relative to the leverage-saliency exerted by other request attributes (i.e., an interaction effect). For example, studies have found that the effects of incentives decrease when topic leverage is high and the topic is made salient. In addition, some have theorized that survey participation, particularly when the sponsoring organization is a major societal institution like the government or local schools, invokes the norms of obligation to one's civic duty. Research has also found that among sample members who have a high level of community involvement, the effectiveness of incentives on survey response is dampened.

The reason for making the previously discussed survey request attributes salient and exerting what leverage they may have on the decision-making process is to counteract the weight of the perceived burden of the survey. If there is one survey request attribute known to all survey researchers to be a burden, it is sample members' perception that completing this survey is a burden on their time. The extent to which survey researchers can reduce the saliency of this perceived burden and increase the saliency of other attributes may extract enough positive leverage to tip the scales in favor of choosing to respond.

Implications for Survey Research

Researchers have tested the theoretical positions of the leverage-saliency theory and have found it to be a useful perspective for describing the behavior of sample members. From these empirical studies of the theory, several implications for survey research can be deduced.

First, sample members have many different reasons for accepting or declining a survey request, and these reasons are often unknown to the person making the survey request. The exception is when an interviewer learns, during introductory contact with a sampled respondent, what the particular respondent is concerned about and interested in.

Second, no one scripted introduction will meet the diverse needs and concerns held by a diverse sample of respondents as it is worded by the researchers. Thus, survey researchers must devise ways of learning to make salient those request attributes that provide the most positive leverage in the decision-making process of the individuals who have been sampled. Research has suggested that expert survey interviewers who are able to engage the sample person in an extended conversation, even if the conversation is not directly related to the survey request, are better able to gauge the sample member's potential concerns and interests by carefully observing the respondent's verbal and nonverbal behavior. Interviewers can then use this information to inform possible hypotheses as to what the concerns of the sample members are and thereby better consider how to tailor their verbal behavior to address these concerns. Through the act of tailoring the survey request, interviewers can be intentional in making salient those survey attributes that may allay the concerns held by the sampled person. Past research has shown that many interviewers can be trained to engage in these behaviors more effectively than what comes naturally to them without the training. This approach to interviewing is sometimes called refusal avoidance training.

The use of tailoring a survey request to address the potential concerns of a sample person is demonstrated as follows. If a survey interviewer approaches a sampled person with a request to complete a survey focusing on illegal drug behavior, the survey interviewer may attempt to gauge how important privacy is to the sampled person and to then make salient the privacy features of the survey request (e.g., the use of computer-assisted self-interviewing equipment). An in-person survey interviewer may pick up on the sampled person's cues regarding personal privacy by noticing if the curtains to the home are open to the street, if the person checks to see who is at the door before opening it, and if the person appears guarded in the presence of the survey interviewer. These may be clues used by the survey interviewer to conclude that privacy holds great leverage by the sampled person and that making the privacy precaution attribute of the survey request more salient will increase the likelihood of the sampled person to accede to the survey request.

A final consideration for the application of leverage-saliency theory applies to panel designs, whereby the same respondents are contacted over time to participate in the survey for more than one wave of data collection. Experience with the respondent from previous waves, including both contact information gleaned from call or contact history data and answers the respondent provided to previous questionnaires, could be used to help tailor the next survey request so as to make it especially salient to the targeted respondent.

Tricia Seifert

See also Interaction Effect; Nonresponse; Refusal Avoidance Training (RAT); Respondent Burden; Response Propensity; Saliency; Tailoring; Unit Nonresponse

Further Readings

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *68*, 2–31.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, *64*, 299–308.

Likely Voter

A likely voter is someone who is registered to vote in an upcoming election and is deemed likely to vote in that election, for example, by pollsters trying to forecast the election outcome. Pre-election pollsters face a unique challenge. At their most germane, they seek to sample a population that is unknown and indeed technically unknowable, because it does not and will not exist until on and around Election Day; this population is the voting public. For a survey that seeks to measure the attitudes and intentions of voters in a given election, the only recourse is to estimate this population through a process known as *likely voter modeling*.

There are no fixed rules for likely voter modeling, and techniques vary. But all begin with a similar approach, using known or self-reported information, or both, about respondents—for example, actual or self-reported voter registration status, actual or self-reported voting history, self-reported interest in the campaign, and self-reported intention to vote—to determine their likelihood of actually participating in the coming election. More controversially, some also may use weighting adjustments for political party affiliation or for demographic targets drawn from exit polls or other sources.

There are three main types of likely voter modeling: (1) screening, (2) scaling, and (3) probability (propensity) modeling. In screening, respondents are identified as likely voters on the basis of their answers to a series of questions. In scaling, or cutoff modeling, qualification requires selected answers to, say, any five of eight questions. The third approach employs a probability model to build a probable electorate in which each respondent is assigned a weight (which can range from 0 to 1) reflecting her or his estimated likelihood of voting.

In the first two approaches, respondents are either classified as "likely voters" and included in the sample, or as "nonvoters" and excluded; in the third, all respondents are included, but with varying weights. Results are identified as representing the views of "likely" or "probable" voters and, in some cases, are distilled further to "certain" or "definite" voters.

Some polling organizations use a single, pre-established likely voter model; others run several models, assessing results across a range of scenarios positing differing levels of voter turnout and then investigating differences across models when they occur. To some extent, all likely voter models involve human (professional) judgment as to the elements they include, the turnout level or levels they anticipate, and the weights applied; at the same time, they are empirically based and ultimately tested (and ideally are later refined) against the actual election outcome.

Likely voter modeling is fraught with hazard. As easily as estimates are improved by good modeling, they can be worsened by poor modeling, for example, through the inadvertent inclusion of nonvoters, the exclusion of actual voters, or both. Poor likely voter modeling is the likeliest cause of inaccurate final estimates in otherwise rigorous pre-election polls.

Poor modeling can negatively impact results well before the final estimate. Ill-conceived likely voter models can introduce volatility in estimates—swings in candidate support that do not reflect actual changes in opinion but rather changes in the characteristics of respondents moving into and out of the model. The goal of good likely voter modeling is to report real changes, not changes that are an artifact of the model itself.

Likely voter modeling increases survey expense (or decreases effective sample size) because it requires discarding or weighting down interviews with nonvoters. To avoid this downside, while still claiming to produce a likely voter survey, some pollsters use weak or lightly screened models that include an unreasonable number of nonvoters. Weeks or months before Election Day, these estimates cannot be held to account by actual results, but they can produce different estimates in different surveys, making variations in models look like volatility in the electorate.

Indeed, one useful way of evaluating a likely voter model is to compare the turnout level it estimates with reasonable expectations for that election. For example, a model that includes 55% of the general population as likely voters in a primary election where anticipated actual turnout is 15% would be a poor one. However, even models that winnow down to an appropriate turnout level may miss the mark by misstating the size of key population groups within the electorate (e.g., senior citizens, African Americans, or women) or the voter preferences within these groups.

In random-digit dialing sampling, a best-practice approach is to interview unlikely voters to obtain sample-balancing demographics to use in weighting, enabling likely voters to be calculated as a subset of the full census-weighted data set. In registration-based sampling, adequate full-population weighting parameters generally are not available. Those data may be unweighted, or weighted to other data, commonly exit poll results from previous, supposedly comparable elections. Estimates from surveys weighted in this fashion can be vulnerable to differential turnout among population groups from election to election. On the other hand, adjusting likely voter data (either obtained via random-digit dialing, registration-based sampling, or a combination of the two) to an empirically based estimate of political party allegiance may compensate for sampling "noise" expressed as trendless night-to-night variability in party identification.

Polling organizations tend to switch from general population or registered voter samples to likely voter modeling over the course of the last few months before Election Day, in an effort to sharpen their estimates by limiting their samples to the most relevant population that can be identified. These polls ask respondents whom they would support "if the election were held today." The election, of course, is not held that day, and that day's likely voters may or may not participate as actual voters on the real Election Day. Nonetheless, within the confines of what these polls attempt to measure, likely voter modeling is a necessary effort—and a largely successful one, given the usual accuracy of final estimates in good-quality pre-election polls.

Gary Langer and Daniel M. Merkle

See also Exit Polls; Pre-Election Polls; Probable Electorate; Random-Digit Dialing (RDD); Registration-Based Sampling (RBS)

Further Readings

Daves, R. (2000). Who will vote? Ascertaining the likelihood to vote and modeling a probable electorate in pre-election polls. In P. J. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 206–223). New York: Chatham House/Seven Bridges Press.

Erikson, R. S., Panagopoulos, C., & Wlezien, C. (2004). Likely (and unlikely) voters and the assessment of

campaign dynamics. *Public Opinion Quarterly*, *68*, 588–601.
Freedman, P., & Goldstein, K. (1996). Building a probable electorate from pre-election polls: A two-stage approach. *Public Opinion Quarterly*, *60*, 574–587.
Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Likert Scale

The Likert scale, named for Rensis Likert (pronounced "lick-urt") who published a seminal report describing its use, possibly is the most widely employed form of attitude measurement in survey research. Similar to nearly all psychometric scale measures, the Likert scale consists of multiple items that typically are summed or averaged to produce a more reliable measure than could be obtained by use of a single item.

The Likert scale is a special type of the more general class of *summated rating scales* constructed from multiple *ordered-category rating items*. Its distinguishing characteristics are as follows:

- Each item uses a set of symmetrically balanced bipolar response categories indicating varying levels of agreement or disagreement with a specific stimulus statement expressing an attitude or opinion (e.g., *Ripe cherries are delicious*).
- The response category points for each item are individually labeled (e.g., *Strongly Agree, Agree, Disagree, Strongly Disagree*).
- The descriptive text of these labels is chosen so that gradations between each pair of consecutive points seem similar.

This sense of equidistance often is reinforced by a set of consecutive integers (e.g., 1, 2, 3, 4) used to label each alternative along the continuum of choices. Although in technical terms, the response format used for Likert scale items is at the ordinal level of measurement, researchers traditionally have used parametric statistics (which assume at least an interval level of data) to analyze Likert scales.

It is commonplace for researchers to use the term *Likert scale* incorrectly. The term often is used to refer to an individual item or the response choice set featured by the items. These usages are improper. Rather, a Likert scale is a multi-item measure, each item in a Likert scale is known as a *Likert item*, and the response categories of a Likert item are known as a *Likert response set*.

Likert response sets may include four or more points, though five categories are traditional. Typical wording labels for the five categories are *Strongly Agree, Agree, Neither Agree Nor Disagree, Disagree*, and *Strongly Disagree*, though certainly other descriptors indicating varying levels of agreement and disagreement are used. Though a five-category set is most frequently employed, many psychometricians advocate using response sets of seven, nine, or even eleven points. Others prefer an even number of response choices, eliminating the neutral alternative to force a positive or negative expression of attitude. Use of four or six categories is the norm when an even number is offered.

Choosing the number of points featured by Likert items should not be driven by personal preference or simply whether one judges it appropriate to prevent fence-sitting by using a forced-choice response set with an even number of response options. Rather, the reliability of Likert scales can be affected by the number of response categories. In situations where low scale score variability is expected, reliability generally can be improved by employing greater numbers of categories. In contrast, when opinion toward the topic is widely divided, scale reliability is largely independent of the number of categories.

Although true Likert items use a symmetrically balanced agree–disagree continuum for individually labeled response choices, several other types of ordered-category items are often and incorrectly referred to as Likert scales or Likert items.

In Table 1, examples A and B are true Likert items; example C is not. More accurately, example C is an ordered-category rating item with the Likert-like properties of a bipolar and balanced response set (reinforced with sequential numbers). It is not a Likert item because the respondent does not indicate degree of agreement with a stimulus attitude statement; rather, the respondent is asked to indicate a level of satisfaction or dissatisfaction with a characteristic of some object or entity. Likewise, neither example D nor example E is a Likert item. Though the descriptive labels in example D are bipolar, they are not constructed using negation to achieve balance, and they do not ask for agreement with a stimulus attitude. In example E, the response choices are neither bipolar nor balanced and do not refer to a stimulus attitude statement. Each is an example of a generic ordered-category

Table 1 Likert versus Likert-like items

A. I think the president has been doing a wonderful job while in office.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
☐	☐	☐	☐	☐
1	2	3	4	5

B. I feel safe walking alone in my neighborhood at night.

Completely Agree	Mostly Agree	Somewhat Agree	Somewhat Disagree	Mostly Disagree	Completely Disagree
☐	☐	☐	☐	☐	☐

C. How satisfied or dissatisfied are you with the reliability of this product?

Very Satisfied	Somewhat Satisfied	Neither Satisfied nor Dissatisfied	Somewhat Dissatisfied	Very Dissatisfied
☐	☐	☐	☐	☐
1	2	3	4	5

D. Compared with adults in general, how would you rate your own health?

Excellent	Very Good	Good	Fair	Poor	Very Poor
☐	☐	☐	☐	☐	☐
1	2	3	4	5	6

E. When you drink coffee, how frequently do you choose to drink decaf coffee?

Never	Rarely	Occasionally	Often	Nearly Always
☐	☐	☐	☐	☐
1	2	3	4	5

F. Choose the one box along the continuum between each pair of antonyms that best describes how you view the service representative who assisted you.

Rude	1	2	3	4	5	6	7	Polite
Intelligent	1	2	3	4	5	6	7	Stupid

rating item. Similarly, neither item in the set shown as example F is a Likert item. Although the response choices are bipolar, they are not individually labeled and no attitude statement is referenced for agreement or disagreement. Rather, each item is known as a *semantic differential* or, especially when the sequenced numeral reinforcement is omitted, a *discrete visual analog scale* item.

Likert items and their corresponding Likert scales are subject to response distortions. Central tendency bias may result from respondent reluctance to select extreme response categories. Acquiescence bias can result when respondents choose to agree with statements because of a desire to choose what they believe is the "correct" or otherwise most appropriate answer. Similarly, social desirability bias may result when respondents attempt to portray themselves or an organization to which they belong in a favorable light. Researchers typically try to attenuate these effects by varying attitude statements so that some are expressed in a positive form while others are expressed in the negative.

After data collection, Likert items may be analyzed individually, or the value scores observed among theoretically related items may be summed or averaged to create a Likert scale. Though it is common to treat a single Likert item's observations as interval level data, it is more conservative to view such data as ordinal; the assumption that all respondents perceive differences between adjacent levels as equidistant is a tenuous one, particularly when smaller numbers of response choices are involved. When treated as ordinal data, Likert item results can be analyzed using nonparametric tests or chi-square tests of association. Likert scale data may be treated as interval data measuring a latent variable, and if relevant assumptions are met, parametric statistical tests (e.g., analysis of variance) can be applied. Finally, data from Likert items or scales sometimes are reduced to the nominal level by combining all agree and disagree responses into one of two corresponding categories (e.g., "accept" vs. "reject"). When this transformation occurs, a chi-square or McNemar test is the statistical procedure typically applied.

Jonathan E. Brill

See also Acquiescence Response Bias; Attitude Measurement; Balanced Question; Bipolar Scale; Forced Choice; Interval Measure; Ordinal Measure; Semantic Differential Technique; Social Desirability

Further Readings

Likert, R. (1932, June). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Masters, J. R. (1974, March). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, *11*(1), 49–53.

List-Assisted Sampling

List-assisted sampling is a technique used in telephone surveys, which utilizes information from the Bell Core Research (BCR) telephone frame and directory listings to produce a simple random sample. This is accomplished by stratifying the BCR telephone frame into two strata. The high-density stratum consists of 100-banks that contain at least one listed number, and the low-density stratum consists of hundreds of banks without a listed number. The proportion of the sample drawn from each stratum depends on the requirements of the study. This technique started to be widely used by telephone survey researchers in the early 1990s because it increased the efficiency of traditional random-digit dialing (RDD) methods, in particular, the Mitofsky-Waksberg method. List-assisted sampling helps to provide a solid foundation, as well as lending statistical justification, for increasing the efficiency of the sample while not sacrificing coverage. As a result, this sampling technique is used widely by telephone researchers to reduce costs and shorten the data collection period. List-assisted sampling can be done in a few different ways, namely, by (a) dual frame design, (b) directory-based stratification, and (c) directory-based truncation. There are some slight biases associated with list-assisted samples in that those 100-banks without a listed number will contain some residential numbers but are not as likely to be included in the sample. However, these biases are minor when compared to other samples with a more complete sample frame, especially when the gains in efficiency are taken into account.

Impetus for a New Method

In the United States, only 20% of all possible telephone numbers are assigned to a residence. This produces problems for researchers conducting surveys with the U.S. population via telephone in that the amount of work that would be needed for interviewers to cull through all the phone numbers in the country is enormous. A telephone sample with only 20% of the numbers reaching the targeted household is extremely inefficient, and as a result, the survey costs increase as does the length of the field period. There have been various sampling methods used in the past to address this problem, but by far the best known is the Mitofsky-Waksberg method. Introduced in the late 1970s, the Mitofsky-Waksberg method of RDD takes advantage of the fact that residential numbers are often clustered together consecutively in the BCR telephone database, which contains information on telephone exchanges and their geographical mapping throughout the United States. By drawing a sample of 100-banks, which consists of the area code, the prefix, and first two digits of the four-digit suffix, and then dialing a randomly selected number within these banks to determine whether residences are contained within the bank, the Mitofsky-Waksberg method culls the nonresidential sample at a much more efficient rate. A 100-bank will only be retained if the random number that is dialed is indeed a residential number; otherwise, the 100-bank is discarded. Once the 100-banks are chosen, then the telephone numbers are generated by assigning a random two-digit number to the end of the 100-bank exchange.

The two-stage RDD sample design is not without its problems, however. First, some clusters of 100-banks may not contain the minimum number (k) of residential numbers required for that bank. Hence, this greatly slows down the efficiency of calling, as all numbers in this bank must be called in order to meet the minimum number of phone numbers. Second, determining the residential status of a number by simply dialing that number is not necessarily a foolproof method. Often the status of the number will be unknown, and the cluster may be rejected erroneously during the first stage of the sample design. Also, the person responding to the initial phone call may regard the number as a residential number, when in reality it may be something other than a residential number, which would then mistakenly make the 100-bank eligible for inclusion. Third, each cluster must be monitored throughout the field period to ensure that k numbers are sampled from the cluster. This is a great drain on resources and often results in longer field periods as cluster yields may only become apparent later on in the data collection period. Further, numbers used as replacements for nonresidential numbers within a given cluster will not receive as many chances for resolution as those numbers identified as residential numbers early on in the field period. Lastly, the most cumbersome problem with the Mitofsky-Waksberg method is the two-stage cluster design it utilizes, which increases the variance of the estimates when compared to a simple random or stratified design. As these problems made themselves more apparent over the course of experience with telephone surveying in the 1980s, a new method was sought.

Directory-Based Sampling

With the rise of national directories of the listed numbers in the United States (e.g., White Pages, MetroMail), calling efficiency was greatly increased. Sample designs could be based on a one-stage selection procedure from these national directories, and only a very small percentage of numbers would be found to be nonresidential, depending on how often the database was updated. While this provided researchers with a purely residential sample, it also excluded numbers that were unlisted residences. This increased the problem of coverage error in telephone surveys. This presented a problem to survey researchers as they realized that not only were they excluding these unlisted residential numbers from the sample frame, but households vary significantly based on their decision to list their phone number. This pressing problem gave rise to list-assisted sampling, which sought to preserve the efficiency associated with the directory-based sampling but also to increase the amount of coverage to something close to the Mitofsky-Waksberg method.

List-Assisted Sampling

To produce a sample frame that utilizes the best in coverage from the Mitofsky-Waksberg method while maintaining the efficiency seen in directory-based sampling methods, the sample frame information used in both of these methods needs to be combined. By enumerating the numbers of the BCR frame and matching these numbers to the directory of listed numbers, researchers are able to establish the listed status of each number on the BCR frame without dialing a single digit. Once the listing status of each number is known, researchers can then draw a sample of numbers directly from the BCR frame, without utilizing a two-stage design. There are three predominant ways of producing a list-assisted sample frame: (1) the dual frame design, (2) directory-based stratification, and (3) directory-based truncation.

Dual Frame Design

In the dual frame design a RDD sample is selected from the BCR frame, and a sample of telephone numbers is selected from the directory frame. These frames are then fielded independently of each other. The efficiencies gained by using a directory-based sampling frame are balanced with the coverage offered by the BCR frame. However, there are problems with this approach. First, the BCR frame still contains all of the problems that were associated with the Mitofsky-Waksberg method: unknown residential status, clustering, inclusion of empty hundreds blocks, and so on. Second, combining the two samples into one data set provides a whole new set of estimation problems. Ideally one would use a one-stage sampling procedure based on a single frame.

Directory-Based Stratification

The directory-based stratification method is a one-stage sample based on the BCR frame with the listed status of each number obtained by comparing the frame to the directory. Once the listed status of each phone number is known, the BCR frame is separated into two strata. The first stratum contains 100-banks with at least one listed number, known as the high-density stratum, and the second stratum contains 100-banks without a listed number, known as the low-density stratum (in this case *density* refers to residential status, not listed status). Phone numbers are then randomly selected from both strata; however, more numbers are usually drawn from the high-density stratum to increase the efficiency of the calling effort. Ultimately, it is up to the researcher, based on the needs of the data collection effort, which percentage of the final sample is drawn from the high-density stratum versus the low-density stratum. Again, the give-and-take between efficiency and coverage play a key role in the decision-making process.

Directory-Based Truncation

Another way to implement a list-assisted sample is to use the truncated design. After the banks are divided into high- and low-density strata, the low-density stratum is dropped entirely. This dramatically improves the efficiency of the dialing; however, the coverage error is increased as a result of the loss of some residential numbers in the low-density stratum. It has been estimated that 55% of all numbers contained in a high-density stratum are residential, given the tendency of residential numbers to be clustered together. However, only 3% to 4% of telephone numbers in the low-density stratum are residential numbers because there are far fewer 100-banks that contain residential numbers but no listed numbers. Hence the bias associated with directory-based truncated designs is only slightly higher when compared to designs that draw sample from both strata.

Bias in List-Assisted Samples

List-assisted samples overcome many of the problems they were designed to conquer. They are much more efficient than traditional RDD methods, while still maintaining the necessary coverage to produce a truly representative sample. In addition, these samples consist of one-stage sampling, which avoids the problems associated with the clustering of 100-banks and assures that all 100-banks have a chance to be selected into the sample frame, albeit some are more likely to be included than others. However, researchers have shown that some bias still is present in list-assisted samples.

By comparing a list-assisted sample using a truncated design to a pure EPSEM (equal probability of selection method) sample drawn from the BCR frame, small differences between the two samples become apparent. When these two samples are compared, no statistically significant differences in measures of sex, race, age, or geographic region emerge. There is a slight difference in education level, where those with higher education are found to more likely be contained in the high-density stratum; however, this is not significant either. Estimates of urban/rural designation are also slightly different, with households in rural areas underrepresented in the high-density stratum. It should also be noted that people who move often will likely be assigned telephone numbers in newly opened 100-banks and therefore will be more likely to be in the low-density stratum due to the lag time in updating the directory listing.

Implications

List-assisted sampling has greatly improved the efficiency of field work with regard to large, random-selection telephone surveys. By incorporating the directory listings into the BCR frame, researchers have found that the problems associated with each frame individually can be overcome when these two frames are combined. No longer beholden to the Mitofsky-Waksberg method of RDD sampling, telephone surveys have been fielded much more quickly (and at less cost), and as a result, study findings have been released to the public (or client) earlier. Increasing efficiency in telephone surveys is not just a matter of saving money but also of maintaining relevance in a fast-paced research world. Undoubtedly the next innovation in telephone sample design will prove to be even more efficient in the future.

Paul Schroeder

See also Coverage Error; Directory Sampling; Dual-Frame Sampling; EPSEM Sample; Listed Number; Mitofsky-Waksberg Sampling; Random-Digit Dialing (RDD); Sampling Frame; Simple Random Sample; Strata; Telephone Surveys; Unlisted Household; Zero-Number Banks

Further Readings

Brick, J. M., Judkins, D., Montaquila, J., & Morganstein, D. (2002). Two-phase list-assisted RDD sampling. *Journal of Official Statistics*, *18*, 203–215.

Brick, J. M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, *59*, 218–235.

Levy P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.

Tucker, C., Casady, R. J., & Lepkowski, J. (1993). A hierarchy of list-assisted stratified telephone sample design options. *Proceedings from the Section on Survey Research Methods* (pp. 982–987). Alexandria, VA: American Statistical Association.

Tucker, C., Lepkowski, J., & Piekarski, L. (2001, May 18). *List assisted sampling: The effect of telephone system changes on design*. Paper presented at the Annual Conference of the American Association for Public Opinion, Montreal, Quebec.

Tucker, C., Lepkowski, J., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly*, *66*, 321–338.

Listed Number

A listed telephone number is one that can be retrieved from a telephone company's directory assistance service, and it also may be published in a local telephone directory. The majority of U.S. residential landline (wired) telephone numbers are listed, but a growing proportion are unlisted—more than 40% as of 2008. There is no equivalent concept of "listing" that applies to cell phone (wireless) numbers in the United States, as there are no directory assistance services or telephone directories that are publicly accessible that contain cell phone numbers.

Whether or not a landline telephone number is listed is predictive of the likelihood that a completed interview will be attained with that household in a telephone survey. A greater proportion of interviews are completed with numbers that are listed than are completed with unlisted numbers. There are two primary reasons for this. First, those people who list their telephone number, in general, are more positively disposed to cooperate when they are reached for a telephone survey, compared to those with unlisted numbers. Concern about privacy is one of the factors that explain this. The demographic correlates of whether or not someone lists their number also are related to this, as minorities (who generally have lower cooperation rates in surveys) are less likely than are whites to have a listed landline number.

Second, almost all listed numbers also have an address associated with them. As such, researchers can send advance mailings to these households when they are sampled for a telephone survey to alert them ("warm them up") to the fact that an interviewer will be calling them. Advance letters with as small an incentive as $2 have been found to raise cooperation rates by approximately 10 percentage points in general population telephone surveys in the United States.

On average, listed telephone numbers require fewer callbacks than unlisted numbers for them to reach a proper final disposition. Thus, the calling rules used by a survey center to process listed numbers should differ from the rules used to process unlisted numbers. However, unless a survey center has their telephone samples screened for listed versus unlisted status or received this information for each number in their sample from their sample vendor, it will not be possible for them to take the listed versus unlisted status into account as their computer-assisted telephone interviewing system processes the callback attempts.

Paul J. Lavrakas

See also Advance Contact; Advance Letter; Calling Rules; Cold Call; Computer-Assisted Telephone Interviewing (CATI); Matched Number; Random-Digit Dialing (RDD); Telephone Surveys

List-Experiment Technique

The list-experiment technique is a survey measurement technique that uses an experimental design to measure a sensitive topic in a way that circumvents much of the questionnaire-related and respondent-related measurement errors that may result from using other methods due to the biasing effects of social desirability and privacy concerns. For example, in 2007, a study using the list-experiment technique reported that more than one fourth of Americans (26%) would be "upset or angry" if the country were to elect a female president. This percentage was much higher than what had been reported previously about the magnitude of this concern from surveys that used more standard (nonexperimental) measurement techniques.

In its simplest form, the list-experiment randomly assigns a sample of respondents to one of two groups: a control group and a treatment group. The control group is exposed to a questionnaire version in which they are asked to consider a list of statements that may or may not apply to them; then they are asked to report "how many" of the statements apply. They are not asked to indicate which of the statements in the list apply, but merely to indicate the total number of statements that apply. For example, if there are four statements, then a control group respondent merely provides an answer from "zero" to "four." In contrast, the treatment group is exposed to a different version of the questionnaire in which they are asked to consider the same statements, but the list they are given includes one more statement than the list given to the control group. Neither the control group nor the treatment group is aware that they are being exposed to a different list. The treatment group is not asked to indicate which statements apply but merely to give the total number of statements that apply. If, for example, the control group is given four statements, then the treatment group is given a list with the same four statements plus one additional (fifth) statement. This additional statement is about the sensitive issue that the researchers are trying to measure accurately.

By comparing the average answer given by the respondents in the treatment group with the average answer given by the control group respondents, the researchers have an unobtrusive way of estimating how many people in the target population the sensitive issue applies to. Because this technique is used when a controversial or otherwise sensitive issue is being measured, there are reasons for the researchers to be concerned if the issue were asked about directly. The concern is that if that were to happen, a reduced proportion of respondents (i.e., biased on the low side) would be identified as having the issue applying to them.

For illustration purposes, imagine that a researcher was interested in measuring whether or not a person had a handgun in his or her home. Asking this question directly in a survey would likely yield biased data because of social desirability and privacy effects. Using the list-experiment technique, the researcher would randomly assign respondents to one of two versions of the questionnaire. One version, the control condition, could contain the list of items shown in Table 1. The other version, the treatment condition, could contain the list of items shown in Table 2. The sensitive item added to the list in Table 2 is *I have a handgun at home*.

The ordering of these items in the two questionnaires should be randomized so that not all respondents are exposed to the same order. In this example, imagine that the researcher found that the respondents in the control group who were exposed to the four statements shown in Table 1 gave an average answer of 1.8, meaning that on average approximately two of the four items applied to them. In contrast, imagine that in this example the respondents in the treatment group who were exposed to the five items gave an average answer of 2.3. The difference between the two group means is $2.3 - 1.8 = 0.5$. Taking this difference (.5) and multiplying it by 100 indicates that 50% of the respondents in the survey have a handgun in their home. This follows from the fact that a controlled experiment was deployed in which the only difference between the two randomly assigned groups of respondents was that the treatment group was

Table 1 Items shown to control group

I went to the grocery store in the past 7 days.
I went to a movie theater in the past 30 days.
I took money out of an ATM yesterday.
I have a high-definition television at home.

Table 2 Items shown to treatment group

I went to the grocery store in the past 7 days.
I went to a movie theater in the past 30 days.
I took money out of an ATM yesterday.
I have a high-definition television at home.
I have a handgun at home.

exposed to one more item in the list than was the control group. Given that this is a design with strong internal (cause-and-effect) validity, the researchers can be confident about the findings.

As powerful as it is in improving the accuracy of measuring sensitive survey topics, the list-experiment technique has a methodological flaw of its own and also has a major analytical disadvantage associated with it. The flaw comes about because anyone in the treatment group who is shown the list with the sensitive issue and for whom all the statements apply is disclosing that fact by giving an answer that equals the total number of items in the list. In the handgun example in Table 2, anyone who says "five" is known to have a handgun. This may cause some people who should answer "five" to instead answer "four" to not allow the interviewers and researchers to know for certain that they have a handgun. Because of this flaw, the list-experiment technique likely yields an underestimate of the prevalence of the sensitive characteristic. A way to reduce the impact of this flaw is for the researchers to make certain that the list of statements shown to everyone contains at least one nonsensitive statement that has an extremely low rate of occurrence. If this is done, there will be very few respondents in the Treatment Group who are put in the position of responding with an answer that gives away that the sensitive statement applies to them.

Analytically, because respondents merely answer with the number of statements that apply, as opposed to indicating which statements apply to them, researchers cannot do analyses on the sensitive item itself at the level of the individual respondent. This is due to the fact that they cannot know whether the individual respondent possesses the sensitive characteristic, except in the case of two types of respondents. In the previously mentioned handgun example, only for people in the treatment condition who answered "five" can it be known for certain that they have a handgun at home and only for those who answered "zero" can it be known they do not have a handgun at home. All others who answered anything other than "five" or "zero" may or may not have a handgun. As such, analyses (e.g., multiple regression) about which of the respondents have handguns, and why, cannot be conducted. Group-level analyses can be conducted, for example, analyses that indicate which demographic characteristics are more likely to correlate with possessing the sensitive characteristic, but not individual-level analyses.

Despite these disadvantages, the list-experiment technique provides researchers with a powerful method to avoid most of the major biasing measurement effects that can and often do result when more direct survey questioning is used to measure sensitive issues.

Paul J. Lavrakas

See also Control Group; Experimental Design; Internal Validity; Privacy; Questionnaire-Related Error; Random Assignment; Randomized Response; Random Order; Random Start; Respondent-Related Error; Social Desirability

Further Readings

Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology*, *22*, 377–399.

Streb, M. J., Burrell, B., Frederick, B., & Genovese, M. A. (2007). Social desirability effects and support for a female American president. *Public Opinion Quarterly*, *72*(1), 76–89.

Webb, E., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures* (Rev. ed.). Thousand Oaks, CA: Sage.

List Sampling

List sampling is one of the basic ways that survey samples can be created. The basic concept of list sampling is deceptively simple. The process is to choose a subset of the elements (the sample) from a listing of all elements (the sampling frame) using a specific selection process. The selection process may have several features, for example, sampling with replacement or sampling without replacement.

In list sampling, as in other sample selection processes, issues arise about whether the sample estimate is an unbiased and reliable estimate for the characteristic or attribute in the full list of elements. Bias and reliability are measures of how well the estimator for the attribute computed using list sample data corresponds to the true value for the attribute in the full list.

Bias is the difference between the true value and the expected value of the estimator. A specific estimator can be determined to be unbiased or nearly unbiased on the basis of sampling theory for the estimator and the sampling process, but the bias cannot be estimated explicitly from a sample.

Reliability is a measure of the reproducibility of the estimate from the sample over repeated application of the sampling process with the same sample size from the same list. The sampling variance is a commonly used measure of an estimate's reliability. Unbiased estimation of the sampling variance requires that (a) every unit in the list has a known, positive chance of being selected (i.e., the unit selection probability is greater than zero) and (b) every pair of units has a positive chance of being in a sample (the joint selection probability for any pair of units is greater than zero).

List sampling can be performed incorporating a number of different sampling processes depending on the analysis objectives for the sample, the information available for each unit in the list, and data collection procedures. As examples, list sampling can be configured as either single-stage or a multi-stage sampling and with or without stratification.

Single-Stage List Sample

The most basic approach to list sampling is an unrestricted simple random sample, which uses a random, equal-probability selection to identify a subset of units on a list for the sample (for a sample size of *n*, all possible combinations of *n* units on the frame have the same chance of being selected). The advantage of this type of list sampling is the ease of use: It can be done using a random number generator on a spreadsheet. The primary disadvantage is that the distributional characteristics of some samples will differ substantially from the distributional characteristics of all the elements in the sampling frame. As an example, a researcher requires an estimate of the percentage of children in a specific State Children Health Insurance Program (SCHIP) who had immunizations in a specific year. The sampling frame is the listing of children served in a specific state by SCHIP during the year. Using an unrestricted simple random sample, a valid random sample may contain only girls or children of a specific age. Although the chances of such samples are extremely small, unrestricted sampling random sampling will allow such samples.

A stratified random sample can avoid such uncharacteristic samples by imposing some control on the sample. Using information on an attribute for each element in the list, this information can be used to define explicit or implicit strata. *Explicit stratification* partitions the full list into two or more mutually exclusive parts; for example, the list of SCHIP children can be partitioned by gender: girls and boys. The number of units selected can be proportionally allocated relative to the number of elements in each stratum, or some of the strata may be assigned a proportionately larger sample size than other strata.

Implicit stratification produces a similar result but does not specify a sample size to each stratum. It requires (a) sorting the sampling frame by the attribute for all elements in the list and (b) the use of a sequential selection process to select the sample. A sequential selection process requires that each element be considered in the order that it occurs in the list and a probabilistic decision is reached for each element concerning the element's inclusion in the sample. With equal probability sequential selection, the sample will be approximately proportionally allocated to each implicit stratum and each sample element would have the same selection probability. For example, implicit stratification with the list of SCHIP children would have the list sorted by gender and then the use of an equal probability sequential selection process to select the sample. The proportion of girls and boys in the sample would closely match the proportion of all girls and boys in the original list.

Explicit and implicit stratification can be used in tandem to control on multiple factors. For example, a list sample of SCHIP children can be explicitly stratified by gender and then implicitly stratified by a second factor (such as age) to further control the characteristics of the sample. Such a sample would contain a specific number of girls and boys. In addition, the age distribution of girls (or boys) within the sample will closely match the age distribution of girls (or boys) in the sampling frame. The advantage of explicit and implicit stratification is that the characteristics of the elements in the sample can be controlled to reflect the distribution of key characteristics of the elements in the sampling frame or, in the case of explicit stratification, to allocate more (or less) sample to individual strata. The disadvantages are that data on the characteristics need to be available for essentially all elements in the sampling frame, and, if the sample is disproportionately allocated to the strata, the sampling variance (computed using the sample weights) for the estimates based on the sample may be increased relative to the sampling variance for an unrestricted simple random sample.

Multi-Stage List Sample

In some surveys, face-to-face interviews may be required or the elements on a list may fall into natural clusters, such as boxes of papers. When face-to-face interviewing is required, it can be cost efficient to form clusters of elements in the same geographic area and assign field interviewing staff to collect the data from all or a sample of elements in the cluster. For example, clusters of SCHIP children can be formed using the zip code or county of residence. A sample of clusters and then all or a sample of the children would be selected.

Explicit and implicit stratification and equal or unequal probability of selection can be used for the selection of both the clusters and the elements within the clusters.

Frank Potter

See also Bias; Sampling Variance; Sequential Sampling; Simple Random Sample; Stratified Sampling

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

LITIGATION SURVEYS

Litigation surveys are surveys that are used in legal proceedings. Like surveys in nonlitigation contexts, they measure opinions, attitudes, and behavior among representative samples of the general public, consumers, employees, and other populations. They afford a useful mechanism for collecting and evaluating opinions or experiences of large groups of individuals bearing on disputed issues in civil and criminal lawsuits. When administered by independent persons who qualify as court expert witnesses in survey methodology by reason of their knowledge, experience, and education, these group assessments may provide more trustworthy information on legal questions than the testimony of a handful of witnesses chosen by an interested party.

State and federal courts throughout the United States now regularly admit survey evidence. However, before 1963, survey respondents' statements were considered *hearsay* (i.e., statements made outside of court, which are offered to establish facts), and such data were disallowed by many American courts. In 1963, the U.S. District Court in *Zippo Manufacturing Co. v. Rogers Imports, Inc*. [216 F. Supp. 670, S.D. N.Y.] recognized that survey evidence fit within certain exceptions to the hearsay rule (i.e., the "expert opinion" exception and the "state of mind" exception), essentially removing the hearsay objection as a bar to the introduction of survey evidence in U.S. courts. Today, surveys also have achieved the status of admissible evidence in legal proceedings in other countries as well, such as Canada, the United Kingdom, France, and Germany.

Litigation surveys provide evidence on a multiplicity of issues and now are used routinely in court as well as for the alternative resolution of civil disputes through mediation and arbitration. Public opinion surveys, consumer surveys, and employee surveys represent the most common forms of surveys for litigation.

Public Opinion Litigation Surveys

In the criminal law context, public opinion surveys are frequently used by defendants in highly publicized criminal cases, such as the Polly Klaas case (e.g., *People v. Richard Allen Davis*), to support motions for change-of-venue due to prejudicial pre-trial publicity. Change-of-venue surveys sample potential jurors in the jurisdiction where a criminal case is pending, measuring their exposure to pre-trial publicity (e.g., whether respondents have heard about the pending matter, and if so, what they have heard) and the potential for prejudice (i.e., whether respondents have formed opinions about the guilt or innocence of the defendant, or "pre-judgments"). To determine whether the potential for prejudice is the same elsewhere, change-of-venue surveys sample from other jurisdictions as well.

In civil litigation, defendants in class action and mass tort lawsuits often conduct public opinion surveys to evaluate their prospects for obtaining favorable jury verdicts and to support change-of-venue motions. Plaintiffs in civil litigation sometimes commission public opinion surveys to determine in which jurisdiction to file a lawsuit where they have the option to select from several jurisdictions.

Consumer Litigation Surveys

Consumer surveys (i.e., surveys of past and potential purchasers of particular products and services) represent the most common form of litigation survey. Because consumer surveys are so routinely used in trademark and advertising controversies, courts actually have commented unfavorably on the failure of a litigant to present survey evidence, and the failure of plaintiffs to provide survey evidence in such matters may give rise to the inference that results of a consumer survey would be unfavorable to their position.

Businesses commonly use trademarks to identify their products and services to consumers and to distinguish them from those of other entities. In trademark disputes, consumer surveys measure whether a name, word, phrase, symbol, design, or a combination of these elements functions as a trademark or whether it is generic (i.e., whether it indicates the source of the product or is a common name for the product itself). Surveys in trademark disputes also assess whether consumers associate a name, word, phrase, symbol, or design with a single source (i.e., whether it has acquired distinctiveness or secondary meaning) and whether consumers are likely to be confused as to the source, sponsor, or affiliation of a product or service because of the similarity of two trademarks.

In copyright disputes, consumers are surveyed to determine whether a product or service (e.g., Napster) is mainly used for infringing purposes (e.g., to obtain copyrighted music) or non-infringing purposes. Additionally, both copyright and patent infringement disputes employ surveys to evaluate whether an infringing feature of a product has influenced consumer purchases and, if so, to apportion damages (e.g., whether the infringing software motivated purchases of a suite of software products).

In advertising disputes, surveys measure consumers' perceptions of what an advertisement or commercial communicates. Consumer surveys are used to evaluate whether an advertisement or commercial has a tendency to create a false or misleading impression among its target audience and, if so, whether the false or misleading impressions are likely to influence purchase decisions.

Consumer surveys also find application to civil actions involving other issues, including rights of publicity, defamation, and product liability.

Employee Litigation Surveys

Employee surveys are becoming the norm in class action lawsuits pertaining to wage and hour issues. Many wage-and-hour class action lawsuits concern whether employees in certain jobs (e.g., assistant store managers, department managers) are properly classified as "exempt" under federal and state law. In the United States, employers classify their employees as "exempt" or "nonexempt" for purposes of complying with the federal Fair Labor Standards Act and various state laws that govern wages and hours. Nonexempt employees (e.g., hourly employees) are entitled to a minimum wage and overtime pay, but employees who are exempted from the Fair Labor Standards Act (e.g., some salaried professionals, executives, administrators, and outside salespersons) ordinarily are not entitled to these benefits or protections.

To determine whether employers have properly classified their employees as exempt (e.g., under the executive and administrative exemptions), exempt employees are surveyed about their job duties, their involvement in various types of decisions (e.g., hiring, firing), their own supervision, and their supervision of other employees. Other wage-and-hour class action lawsuits concern whether employers have complied with the Fair Labor Standards Act and other state laws governing their nonexempt employees. For these cases, nonexempt employees may be asked questions regarding off-the-clock work, time record accuracy, meal breaks, and rest breaks.

Employee surveys have shed light on alleged violations of the Equal Pay Act of 1963 and Title VII of the Civil Rights Act of 1964, which require that men and women be given equal pay for equal work and protect individuals against employment discrimination on the basis of sex. Employee surveys in gender discrimination cases collect information regarding the qualifications, experiences, and preferences of male and female employees. This information is used to assess whether pay and promotion disparities are due to impermissible corporate policies and practices or other factors.

Surveys as Scientific Evidence

In the United States, for survey evidence to be admissible, trial courts must determine whether it is relevant, as well as reliable and trustworthy. For such determinations, trial courts now employ the scientific validity test that was initially pronounced by the U.S. Supreme Court in 1993 in *Daubert v. Merrell Dow Pharmaceuticals, Inc*. [509 U.S. 579]. In nonlitigation contexts, the usefulness of a survey also is a function of its applicability, reliability, and validity. Thus, survey methodology standards and principles used for conducting and evaluating litigation surveys essentially are the same as those used in nonlitigation contexts. General survey standards can be found in codes of conduct and guidelines published by professional associations in the survey industry (e.g., American Association for Public Opinion Research, Council of American Survey Research Organizations). Because courts have recognized the probative value and trustworthiness of survey evidence, its application to legal questions will likely continue to develop and expand in the future.

E. Deborah Jay

See also American Association for Public Opinion Research (AAPOR); Council of American Survey Research Organizations (CASRO); Reliability; Survey; Validity

Further Readings

Diamond, S. S. (2005). Reference guide on survey research. In M. J. Saks, D. L. Faigman, D. H. Kaye, & J. Sanders (Eds.), *Annotated reference manual on scientific evidence* (2nd ed., pp. 353–399). Eagan, MN: Thomson West.

Herr, D. F. (2006). *Annotated manual for complex litigation* (4th ed.). Eagan, MN: Thomson West.

McCarthy, J. T. (2006). *McCarthy on trademarks and unfair competition* (4th ed.). Eagan, MN: Thomson West.

National Jury Project. (2002). *Jurywork: Systematic techniques* (2nd ed.). Eagan, MN: Thomson West.

Log-In Polls

A log-in poll is an unscientific poll that typically is conducted by news and entertainment media on their Web sites to engage their visitors (audiences) by providing them an opportunity to register their opinion about some topic that the media organization believes has current news or entertainment value. Typically two choices are given for someone to express her or his opinion. One choice might be for those who agree with the issue and the other might be for those who disagree. For example, a log-in poll question might be to indicate whether a Web site visitor agrees or disagrees that *Congress should impeach the President*.

These polls are not accurate measures of public opinion on the topic. The people who choose to register their opinion on the Web site represent no known target population, and as such, the media organization cannot know to whom the findings generalize. Most often, response options such as "undecided" are not given as choices. This further invalidates the data, as anyone who holds an opinion other than what the limited response choices reflect or holds no opinion on the topic (e.g., has no previous knowledge of the topic) cannot indicate that on the poll. Furthermore, oftentimes the Web site allows people to register their opinion more than once, by exiting the Web site and then returning at another time, which is a further indication of the unreliable data these pseudo-polls generate. In fact, the data these polls generate do not even necessarily represent the opinions of visitors to the Web site since the method of sampling is self-selection, which is unreliable. Although log-in polls may provide some entertainment value for the media organization and its visitors, especially those who register their opinion, they should not be considered accurate news.

A real danger these pseudo-polls can create is when their results are disseminated uncritically as though they had some degree of validity. For example, a talk-radio program host might cite the "latest results" from the station's Web site poll to stimulate conversation about the issue from call-in listeners. As the on-the-air conversation about the poll topic proceeds, new listeners who were not tuned-in when the log-in poll results were first mentioned may have no sense how the topic of discussion started. Additionally, some listeners may assume the topic is a controversial one among large portions of the public, and the topic may get further discussed and disseminated via other uncritical news outlets and by word of mouth. Thus, what started out as a pseudo-poll finding with no basis of scientific support can take on a life of its own far beyond what it should have received.

Log-in polls also are used by organizations other than the media to gather information about visitors to their Web site. But again, the participants are self-selected, and the data are unlikely to provide the organization with reliable information about their population of visitors.

Paul J. Lavrakas

See also Call-In Polls; Computerized-Response Audience Polling (CRAP); 800 Poll; 900 Poll; Pseudo-Polls; Self-Selected Listener Opinion Poll (SLOP)

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Longitudinal Studies

Longitudinal studies or panel studies are studies where the research settings involve multiple follow-up measurements on a random sample of individuals, such as their achievement, performance, behavior, or attitude, over a period of time with logically spaced time points. The purpose of longitudinal research studies is to gather and analyze quantitative data, qualitative data, or both, on growth, change, and development over time. Generally, the significance of the longitudinal research studies stems from the fact that the knowledge, skills, attitudes, perceptions, and behaviors of individual subjects usually develop, grow, and change in essential ways over a period of time. Longitudinal studies require formulating longitudinal research questions and hypotheses, using longitudinal data collection methods (e.g., panel surveys), and using longitudinal data analysis methods.

Researchers across disciplines have used different terms to describe the design of the longitudinal studies that involve repeatedly observing and measuring the same individual subjects (respondents) over time. Some of the terms used are *longitudinal research designs, repeated-measures designs, within-subjects designs, growth modeling, multi-level growth modeling, time-series models*, and *individual change models*.

Advantages

Compared to the cross-sectional research designs, longitudinal research designs have many significant advantages, including (a) revealing change and growth in an outcome (dependent) variable (e.g., attitude, perception, behavior, employment, mobility, retention), and (b) predicting the long-term effects of growth or change on a particular outcome (dependent) variable. Most importantly, longitudinal research studies can address longitudinal issues and research questions that are impossible to address using the cross-sectional research designs. Across all disciplines and fields of study, with the advancement in technology and the use of high-speed computers, more and more data are being collected over many different occasions and time points on the same individuals, leading to complex longitudinal data structures.

Challenges

Such longitudinal research studies present researchers and evaluators across all disciplines with many methodological and analytical challenges. For example, a common problem in analyzing longitudinal data in many disciplines is that complete data for all measurements taken at different time points for all individuals may not be available for many reasons. One possible reason is that some subjects are not available for some of the data collection time points to provide measurements or responses. Another reason is that some subjects might drop out from the study in any time point, that is, attrition. Further, mortality (attrition) can be another reason for having incomplete longitudinal data to make valid conclusions about growth or change.

Categories

Longitudinal research designs and the corresponding analytic methods can be classified into two broad categories based on the methodological and statistical assumptions of each category.

Traditional Longitudinal Data Analysis

Longitudinal data can be analyzed using repeated-measures analysis via SPSS (Statistical Package for Social Sciences) or SAS (originally "statistical analysis software") software when the individuals' longitudinal repeated measurements on a dependent variable are taken over different periods of time. More complex repeated-measure designs are the ones that have at least one independent between-subjects factor (e.g., gender, grade, ethnicity) in addition to having the individuals' longitudinal repeated measurements on the dependent variable taken over different periods of time (within-subject factor). This type of longitudinal design, with both within-subjects factors (repeated measurements) and between-subjects factors (independent variables), can also be analyzed using factorial repeated-measures designs via SPSS.

Using these traditional repeated-measures analytic methods requires the complete longitudinal data, where every individual has all the measurements for all the time points with equal time intervals between the repeated measurements. Missing longitudinal data for some individuals at different time points or having unequal time intervals between measurements poses great complications for longitudinal researchers in using these traditional statistical methods to analyze the longitudinal data.

Multi-Level Longitudinal Data Analysis

Multi-level longitudinal data analysis methods can be used as alternative analytic methods to the traditional repeated-measures data analysis methods to overcome the incomplete data limitations. The longitudinal data can be analyzed via multi-level modeling procedures with software such as hierarchical linear modeling. Such software is designed to analyze various kinds of multi-level data, including multi-level longitudinal data. In the simplest two-level longitudinal modeling conceptualization, the repeated measurements are viewed as nested within the individual rather than as the same fixed set of measurements for all individuals. Thus, both the number of measurements and the timing of measurements may vary in the multi-level modeling and analysis.

These methods are generally more flexible and powerful than the traditional repeated-measures analysis in terms of handling longitudinal data with missing measurements for one or more time points over the longitudinal data collection period and handling longitudinal data sets with more than two levels. For instance, repeated measurements for each individual (level-1) nested within individuals (level-2) who are then nested within groups (level-3, classrooms, sites, or organizations) is an example of three levels of longitudinal data. Multi-level longitudinal data analysis also allows for the flexibility to include level-1 time varying covariates to the repeated data set (e.g., number of days absent from school or work in each time point) as an explanatory predictor variable.

Sema A. Kalaian and Rafa M. Kasim

See also Attrition; Cross-Sectional Data; Cross-Sectional Survey Design; Panel Data Analysis; Panel Survey; Research Design; Trend Analysis

Further Readings

Kalaian, S. A., & Kasim, R. M. (2006). Longitudinal/repeated measures analysis. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 2, pp. 558–561). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

M

Mail Questionnaire

The term *mail questionnaire* refers to the instrumentation of a self-administered survey that has been laid out and reproduced on a paper-based printed medium with the intention that data collection operations will be implemented via traditional postal service deliveries. A survey researcher's solicitation and collection of data via postal communications that include a mail questionnaire is called a *mail* (or *postal*) *survey*.

In mail surveys, the initial delivery of the mail questionnaire is typically in the form of a survey packet. In addition to the mail questionnaire itself, this packet typically includes a cover letter explaining the purpose of the study and encouraging participation, a postage-paid pre-addressed return envelope, and some form of pre-paid incentive or participation gift intended as a social gesture of appreciation. (A crisp, new, uncirculated $1 bill is perhaps the most commonly used incentive device, particularly for commercially conducted mail surveys.) The survey packet often is preceded by an advance notification of some sort (e.g., a postcard informing the recipient that a research questionnaire will soon follow in the mail), and several reminder communications (which usually include one or more replacement questionnaire mailings to nonresponders) are typical.

Advantages

As a data collection methodology, mail questionnaire research offers several advantages. One advantage is low cost relative to costs for similar-quality surveys using interviewer-administered research methods. At one time, the mail survey was clearly the low-cost king, but the emergence of Internet-based surveys offers researchers a second lower-cost alternative.

Another important benefit of the use of mail questionnaires is that, when properly designed and executed, the data collected are generally of high quality. That is, the psychometric performance claimed for scale measures is typically realized. This is not surprising, since a substantial proportion of measurement scales commonly used in basic social research have been developed using self-administered paper questionnaires. Furthermore, the fact that the mail questionnaire is paper based may conjure up the feeling of an examination, resulting in relatively higher levels of respondent effort and attentiveness in filling out the form; indeed, there is evidence that other forms of self-administered interviewing, such as computer-based Web surveys or kiosk surveys, may not yield data of similar integrity.

A third advantage is that, when professionally and diligently implemented among sample cases that are accurately targeted to the sample population, mail questionnaire surveys can be expected to achieve response rates that are similar to or even higher than those that would be achieved by interviewer-administered methods.

Although many factors that can be manipulated by the researcher have been associated with the achievement of mail survey response rates—for example, type of postage used, content and appearance of the cover letter, type and value of the incentive—the characteristics of the mail questionnaire itself appear to be among the most important determinants. Mail questionnaires that have a clean, professional, and uncluttered appearance generally produce higher survey response rates than those lacking these qualities. Mail questionnaires with good face validity, conveying a sense that the survey represents a well-conceived scientific effort featuring questions that are relevant and salient to the purpose of the research, are more likely to gain cooperation than those that lack this quality. In addition, mail questionnaires that include only (or nearly only) precoded questions, either traditional closed-ended questions featuring specified sets of response choices or self-coding open-ended questions, tend to encourage survey participation. Widespread use of self-composed answer open-ended questions, where the respondent must write out an answer, tends to inhibit survey response and often yields incomplete and otherwise lower-quality data, not to mention difficulties of legibility.

Disadvantages

Despite the substantive and important benefits that can be realized through the postal survey methodology, mail questionnaires also have many drawbacks.

Sampling Control. There is only a moderate level of sampling control. While the researcher is able to identify potential respondents by address, it must be assumed that the intended respondent associated with an address is the individual who actually completes the mail questionnaire interview form.

Contingency Questions. Contingency questions, when a particular answer to one question creates the need for a skip pattern to another question later in the questionnaire, can be problematic. Thus, it is very important that researchers pretest how well skip instructions can be followed by respondents, especially when a skip leads a respondent to jump several pages ahead in the questionnaire.

Corrections and Changes. The pre-printed nature of the survey instrument offers little flexibility to the researcher. Corrections and changes once a survey field period starts, if possible at all, are difficult to implement reliably. Likewise, though experimental design manipulations are possible—for example, different ordering of questions or response alternatives across different randomly assigned respondents—they are challenging and burdensome to implement. Multiple versions of the mail questionnaire instrumentation must be produced, and procedures ensuring accurate tracking and control of mailings must be instituted. At the same time, these control processes are fraught with numerous opportunities for human error, inasmuch as the use of automation to implement such designs is neither practical nor feasible given the relatively small volume of mailing pieces typically involved.

Missing Data. The use of mail questionnaires tends to yield relatively high rates of missing data (item nonresponse) relative to interviewer-administered and those self-administered methods involving computerized technology. Unlike these other methods, mail questionnaires do not provide a practical mechanism to monitor or interactively encourage item response compliance.

Data Entry. Entry of the data recorded by mail questionnaires is laborious and time consuming. While optical scanning and programmable template mapping software applications do exist for speeding the data entry of paper questionnaires, these technologies must be implemented following data collection episodes, adding another step to the research process. Perhaps even more important, the reliabilities of these technologies often have been disappointing. Optical scanning also can restrict the layout (format) of the questionnaire, making it less than user-friendly for the respondent.

Presentation Order. There is no ability to control presentation order of items; respondents can examine the contents of the entire instrument before answering any question.

Jonathan E. Brill

See also Advance Letter; Closed-Ended Question; Contingency Question; Cover Letter; Gestalt Psychology; Graphical Language; Incentives; Mail Survey; Missing Data; Open-Ended Question; Precoded Question; Questionnaire Design; Self-Administered Questionnaire; Visual Communication

Further Readings

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method*. Hoboken, NJ: Wiley.

Mail Survey

A mail survey is one in which the postal service, or another mail delivery service, is used to mail the survey materials to sampled survey addresses. What is mailed usually consists of a cover letter, the survey questionnaire, and other materials, such as a postage-paid return envelope, an informational brochure to help legitimize the survey organization, detailed instructions about how to participate in the survey, and/or a noncontingent cash incentive.

In some mail surveys, it is the household or the business at the address that is sampled, but in other mail surveys it is a specific person at the address who is sampled. In the case of a specific person being sampled, sometimes there is a *specifically named* person (e.g., Martha Johnson) who is sampled and other times it is the person with some specific characteristic, such as "householder" or "Chief Information Officer." In most instances, respondents are asked to mail back the questionnaire to the researchers once they have completed it. Some mail surveys provide respondents with multiple modes to choose for their response, including dialing into a toll-free telephone number or going to an Internet site, to complete the questionnaire, rather than mailing the questionnaire back in the return envelope the researchers provide.

Advantages and Disadvantages

By far, the most common advantage of carrying out a mail survey is the cost. It is relatively low priced compared to telephone and in-person modes of surveying, can be used to survey very large numbers of respondents in relatively short periods of time, and is especially cost-effective if the respondents are dispersed geographically. Oftentimes, data collection for a mail survey can begin more quickly than for a survey that involves interviewers, because of the time required to hire and train interviewers and the programming that is required to computerize and test a final version of the questionnaire that will be used in interviewer-administered surveying.

Another advantage of mail surveys is that respondents are afforded the time to produce answers that might be thought through more carefully, as opposed to when an interviewer-administered survey is conducted. Also, respondents can answer the survey questions at their convenience. Furthermore, the respondents are given privacy, which often is an important factor in their deciding to cooperate and in deciding to provide accurate responses, especially to sensitive questions. Visual and/or audio aids included in the mailed package can assist the respondents in completing the survey process accurately.

Recently, there has been a renewed interest in the United States in mail surveys of the general public as the difficulties of gathering survey data from the general public via telephone and concerns about coverage error have grown. In particular, with the rapid increase of the U.S. cell phone only population, it has become much more difficult for telephone surveys to reach their intended target population using the traditional random-digit dialing (RDD) landline telephone frame. At the same time, surveying persons reached via a cell phone in the United States is a very complex and costly undertaking, with many unknowns and uncertainties existing about how to do it right. In comparison, interest in address-based sampling frames that are wholly appropriate for mail surveys is rapidly increasing.

A major disadvantage of a mail survey is the length of the field period that is required to gain a reasonably high response rate, due to the multiple follow-up mailings this requires. Another disadvantage with mail surveys is that many of their operational components must be carried out manually (e.g., coordinating the processing of incoming returned questionnaires), which makes them more subject to error and takes longer to complete. Another possible disadvantage is that some portion of the sample will not have adequate literacy ability to understand some or all of the questionnaire.

Reducing Error

Three major types of survey error should be guarded against when conducting a mail survey. These are (1) coverage bias, (2) unit nonresponse error, and (3) error due to missing data (item nonresponse error).

Coverage Bias

One area in which problems frequently arise in mail surveys is coverage bias. This can occur when a sample is selected from an incomplete sampling frame. For example, an outdated list will produce outdated samples. Problems may also occur when lists are overstated, that is, have duplicate records.

Unit Nonresponse Error

The largest concern about mail surveys is unit nonresponse. In convincing sampled respondents to cooperate with a mail survey request, there generally is no interviewer involved in recruiting respondents and thus in persuading the reluctant ones to cooperate. As such, almost all mail surveys that strive for reasonably good response rates must do multiple follow-up mailings to initial nonresponders. This adds to expense but is a necessary technique for gaining good response from a sample in a mail survey. Without follow-up mailings, which obviously lengthen the field period of the survey, a likely majority of the sample will not respond at all. If the nonresponders differ from responders in the variables of interest to the researchers, then the larger the nonresponse rate, the greater the extent of nonresponse bias in the survey estimates.

Item Nonresponse

Because a mail survey questionnaire is not administered by an interviewer, there is no one present at the time data are being entered into the questionnaire to persuade a respondent to answer all the questions asked and to answer them fully and accurately. To the extent certain questions are avoided (improperly skipped) by respondents, and if those respondents who do not provide accurate answers differ on the variables from those respondents who do answer these items, then missing data will bias survey estimates for these variables.

Basic Mail Survey Considerations

The first consideration in doing a mail survey is deciding if it is a correct data collection mode for the needs of the research question that is being investigated. If it is, then a mail survey properly conducted can generate a high response rate and high-quality data.

Pre-Notification

Pre-notification is when respondents receive advance contact, usually either by mail or phone, notifying them that they have been selected to be in a mail survey and that they soon will be receiving the survey packet. When advance contact is made via the mail, it typically is made through use of a postcard. Advance contact letters that include a token cash incentive (e.g., $1) have been found to effectively raise mail survey response rates.

Survey Envelope

If there is a stamp on the survey envelope or packet that is mailed to sampled respondents, they are less likely to believe that it is a useless piece of mail, such as junk mail. The physical appearance of the survey envelope also will affect the likelihood of whether the recipient ever opens it. Past research has shown that different demographic cohorts react differently to mail envelope appearances; for example, younger adults are more likely to cooperate when sent a bold eye-catching envelope, whereas middle-aged adults have been found to be more likely to cooperate when sent a more conservative-looking envelope. Whenever a mail survey is sent to a specifically named person, it is advisable to use that person's name on the mailing envelope. However, if the survey is directed to anyone who resides or works at an address that has been sampled, then it is best not to use specific names (e.g., the Jones Household) as there may be new persons at the address whom the researchers actually do want to survey. In this case, some generic identifier such as "Research Household" may better serve the researcher's purposes.

Researchers also should give explicit thought to the order of the materials that are placed inside the outgoing survey envelopes, as the order in which a respondent encounters the materials upon opening the envelope may affect her or his decision to cooperate.

Cover Letter

This letter describes the study's purpose, explains the procedures to be followed, assures confidentiality, and strives to motivate participation. The letter should be written in a precise style and should fit onto one side of letterhead stationery. It should have contact information (e.g., a toll-free number or Internet address) in case the respondent wants to contact the researchers and should grasp the reader's attention in the very first sentence. The aim and content of the letter should be written in language that can be clearly understood by respondents of below-average educational attainment. How the cover letter is signed may also affect mail survey response rates, with real (hand-written) signatures often showing the highest response rates. It also is recommended that the cover letter (and the questionnaire) contain a specific date that serves as a deadline for returning the completed questionnaire. When respondents are presented with a deadline, they seem to try harder to return the questionnaire rather than postponing it. Research has shown that deadlines may not so much affect the final response rate for a mail survey, but they do affect the timing of the returns.

Return Envelope and Return Postage

To increase response rates, the respondent should be sent a return envelope addressed to the researcher. This should be mailed to the respondent along with the cover letter and the questionnaire. Postage-stamped return envelopes have been found to lead to higher response rates than return envelopes that have metered postage affixed.

Confidentiality and Anonymity

If respondents believe that their answers will be kept confidential, they are more likely to cooperate. In order to better maintain confidentiality and reassure the respondent, names and addresses should not be placed on the questionnaires. Instead an ID code that only the researchers can link to a specific respondent should be placed on the questionnaire. Returned questionnaires should be stored in a highly secure location with limited access, and respondents should be assured that this will happen.

If the mail survey results are truly anonymous, different procedures are used than when data are merely kept confidential. With confidentiality, the researchers will know something about who it was that completed a specific questionnaire, whereas with anonymity the researchers do not know anything about who completed the questionnaire. However, many respondents misunderstand the difference between confidentiality and anonymity, and thus mail survey researchers need to think very carefully about whether they want to give up certain control that anonymity requires, such as not being able to limit follow-up mailings only to previous nonresponders.

Reminder and Follow-Up Mailings

Reminder mailings are likely to be the most important technique for producing high response rates in mail surveys. Reminder mailings typically contain a modified version of the cover letter, a new questionnaire, and a new return envelope. Without reminder mailings, nearly all mail surveys will achieve far less than a 50% response rate. The return pattern for a given mail survey is revealed as the researchers tally their daily incoming mail. The vast majority of returns for a given survey mailing will be received by the researchers within 2 weeks (14 days) of the date of the mailing of the original survey packets to respondents. When follow-up mailings are sent to previous nonresponders, the researchers likely will experience a lower reply rate than with the previous mailing unless new incentives are used or other aspects of the survey method are altered. Since each reminder mailing yields diminishing returns for the researchers, experience suggests that the total number of mailings should be limited to a maximum of four from a cost/benefit standpoint.

If the mail survey is anonymous, the researchers must mail reminder mailings to all sampled respondents, since the researchers will not be able to tell which respondents already have replied. This is a costly prospect and also is a nuisance to those respondents who already have cooperated. An alternative to a standard reminder mailing in an anonymous mail survey is simply to mail a reminder postcard.

Incentives

Incentives are the second most important way to improve mail survey response rates. Overall, the primary purpose of an incentive is to provide tangible evidence to the respondent that the researcher appreciates the cooperation the respondent will provide. Past research has shown consistently that a noncontingent incentive included in the original mailing will yield higher response rates than a promised (contingent) incentive of greater value. Cash money is the simplest, most direct, and most effective incentive; in general, the larger the value of the incentive, the greater the response rates, but with diminishing returns. Respondents can be rewarded with other things rather than money; for example, incentives have included ballpoint pens and movie tickets.

Past research has shown that the highest return rates were found when both reminder mailings and noncontingent incentives were used. For a researcher with a short time frame for the data collection, relatively more resources should go into incentives than into reminder mailings. In contrast, if time is not the issue but the budget is, then relatively more resources should go into reminder mailings than into incentives.

Length and Look of the Questionnaire

The research literature provides mixed evidence about whether the length of a mail survey questionnaire affects the amount of cooperation and the quality of data the researchers receive. This is complicated by the effect that the topic of the survey has on the interests of respondents. As with all surveys, a respondent will cooperate more fully, including providing good-quality data to a lengthy questionnaire, if she or he is interested in the topic. The implications of this are that researchers, to the greatest extent possible, should gather all data that are truly necessary for a survey project but also keep in mind the likely interest levels of the respondents when determining what other methods to use to gain high levels of cooperation and compliance from respondents.

Considerations also should be given to factors that can affect the perception a respondent forms about how long the questionnaire will take to complete. These factors include those related to the formatting and printing of the questionnaire. For example, printing questionnaire pages back to back will substantially reduce the total number of pages, but it must be done in a manner that does not create an onerous burden to respondents as they turn from page to page. To achieve a booklet effect, the questionnaire can be printed on 17 × 11-inch paper; that is, two pages of the questionnaire per side of a sheet of paper. Another effective appearance is the use of the two-column newspaper-type format. This effect generally reduces the number of pages for a questionnaire. The questionnaire's pages should be balanced, with a fair amount of white space around each question rather than having the questions compacted together. It is important to have a type size and font that is simple rather than overdone with fancy styling. All questions should have a question number. This can take the form of being sequential from start to finish for the entire questionnaire or sequential within sections. There should also be a common style used for laying out the questions, response alternatives, and the instructions through the pages of the questionnaire.

Managing a Mail Survey

Mail survey management involves attending to two primary areas of concern. One is to manage the schedule of all the mailings and the other is to manage the processing of all incoming returns. To do this well, a good quality-control system should be instituted for both processes.

Oversight of the outgoing mailings includes managing all the printing that must be done and assembling (stuffing) all the envelopes that must be mailed out. As mindless an activity as stuffing envelopes can be, errors due to carelessness can be frequent if the staff members assigned to this task are not attentive to what they are doing. As part of quality control, whoever is overseeing the mailings should randomly sample outgoing envelopes before they are mailed to determine the quality with which they were assembled.

Oversight of how the incoming returns are processed is also extremely important. Each incoming returned envelope must be dated and opened soon after it arrives to determine what exactly was returned. That is, not all returns will be completed questionnaires; some respondents will return blank questionnaires with or without indication of why they did so, and some will contain returned cash incentives. Once it has been determined what was returned in each envelope, this information is logged into the database being used to help track the survey's progress. Essentially this is done on a daily basis for each day that incoming mail is delivered throughout the field period, so that the researchers can receive daily information about whether the sample is performing as expected. In this way, changes to the methodology can be made if necessary, for example, deciding to extend the field period with another follow-up mailing or to change the incentive sent in the next follow-up mailing.

Isaac Dialsingh

See also Address-Based Sampling; Advance Contact; Anonymity; Confidentiality; Contingent Incentives; Coverage Error; Cover Letter; Gestalt Psychology; Graphical Language; Missing Data; Noncontingent Incentives; Quality Control; Questionnaire Design; Questionnaire Length; Sampling Frame; Total Design Method (TDM); Unit Nonresponse

Further Readings

Bourque, L., & Fielder, E. (2002). *How to conduct self-administered and mail surveys*. Thousand Oaks, CA: Sage.

Dillman, D. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.

Fink, A. (2005). *How to conduct surveys: A step by step guide*. Thousand Oaks, CA: Sage.

Mangione, T. W. (1995). *Mail surveys: Improving the quality*. Thousand Oaks, CA: Sage.

Trussell, N., & Lavrakas, P. J. (2004). The influence of incremental increases in token cash incentives on mail survey response. *Public Opinion Quarterly*, *68*(3), 349–367.

MAIN EFFECT

A *main effect* is a statistical term associated with experimental designs and their analysis. In the analysis of variance statistical test, which often is used to analyze data gathered via an experimental design, a main effect is the statistically significant difference between levels of an independent variable (e.g., mode of data collection) on a dependent variable (e.g., respondents' mean amount of missing data), ignoring the influence of other factors. To better understand the statistical concept of a main effect, it is helpful to understand a few key terms and experimental conditions under which a main effect may be found.

When conducting research, it is not uncommon to use a factorial analysis of variance to determine how two or more categorical independent variables (called *factors* in analysis of variance) affect a continuous dependent variable. Each factor in a factorial analysis of variance contains two or more categories or *levels* of that factor that are manipulated to determine how the factor influences the dependent variable.

For example, a survey researcher investigating item nonresponse may want to know how the two factors *survey type* (containing the levels "paper-and-pencil survey" vs. "computer-assisted survey") and *mode of administration* (containing the levels "interviewer-administered" vs. "self-administered") separately and together influence the dependent variable *percentage of item nonresponse*. A sample of respondents is randomly assigned to one of the four conditions in the experiment: (1) paper-and-pencil interviewer-administered, (2) paper-and-pencil self-administered, (3) computer-assisted interviewer-administered, and (4) computer-assisted self-administered. A factorial analysis of variance can be used to investigate the main effects of the two factors (*survey type* and *mode of administration*) on the amount of item nonresponse.

In such a factorial analysis of variance, a main effect is a statistically significant difference between the levels of one factor on the dependent variable regardless of the influence of any other factor. In this survey research example, a main effect for the factor "mode of administration" would occur if self-administration resulted in a statistically significant difference in the average amount of item nonresponse when compared to interviewer administration, regardless of any influence that the factor "survey type" (paper-and-pencil vs. computer-assisted) might have on item nonresponse. Ignoring the influence of all other factors on the dependent variable when determining a main effect is referred to as *collapsing* across levels of the other factor. The illustrations in Figures 1 and 2 help visualize this process.

Figure 1 illustrates a main effect for mode of administration. The two parallel lines on the graph show a difference in the amount of item nonresponse between self-administered surveys and interviewer-administered surveys. Mentally collapsing across the factor *survey type*, one can see that self-administration resulted in more item nonresponse ($\bar{X}_s$) than interviewer

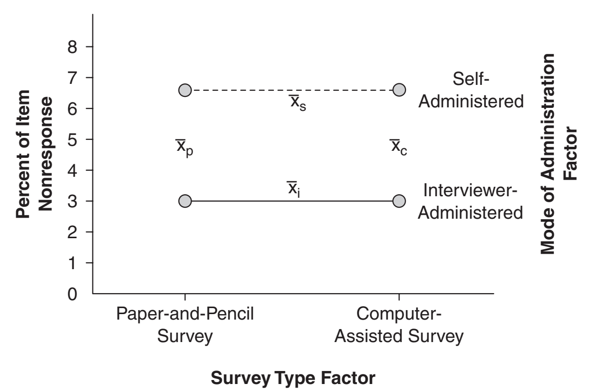

Figure 1 A main effect for mode of administration

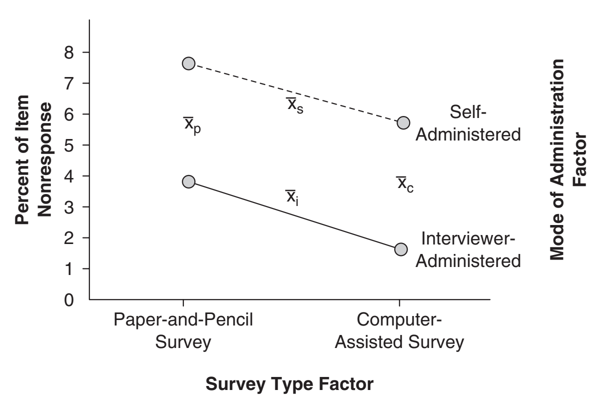

Figure 2 Main effects for survey type and mode of administration

administration ($\bar{X}_i$). There is no main effect for *survey type* because each level of that factor contains identical amounts of item nonresponse ($\bar{X}_p$ and $\bar{X}_c$).

Figure 2 shows a slightly more complex relationship, with main effects for both *survey type* and *mode of administration*. The main effect for the factor *survey type* shows that paper-and-pencil surveys have a greater amount of item nonresponse ($\bar{X}_p$) when collapsed across *mode of administration* than the amount of item nonresponse in computer-assisted surveys ($\bar{X}_c$). The main effect for the factor *mode of administration* shows that self-administration results in more item nonresponse ($\bar{X}_s$) than interviewer administration ($\bar{X}_i$) when collapsed across levels of survey type.

Main effects are not determined by merely eyeballing graphs, however. Identifying main effects requires statistical calculations examining whether the differences between levels of a factor on the dependent variable are no different than would be expected due to chance. The statistical calculation examining main effects results in an F-statistic that is computed by dividing the total variance between levels of the factor by the total variance within the factor. The Statistical Package for the Social Sciences (SPSS) and SAS are two statistical software packages that can compute main effects using factorial analysis of variance and will inform the user whether there are statistically significant main effects.

In addition to main effects, factors can interact with one another to produce an *interaction*. If an interaction is present, caution should be used when interpreting main effects.

Dennis Dew

See also Analysis of Variance (ANOVA); Dependent Variable; Experimental Design; Factorial Design; *F*-Test; Independent Variable; Interaction Effect; Random Assignment; SAS; Statistical Package for the Social Sciences (SPSS)

Further Readings

Hays, W. L. (1994). *Statistics*. Belmont, CA: Wadsworth.

McCall, R. B. (2001). *Fundamental statistics for behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

MALL INTERCEPT SURVEY

The mall intercept survey is used most frequently for market research. The name is derived from its traditional deployment at shopping malls, where there are many readily available persons to sample and from whom to gather data. Mall intercept studies rarely use probability sampling methods, and therefore the results of such a survey cannot be used to determine scientifically the attitudes and opinions of the target population. To address this limitation, researchers often attempt to conduct the survey in a number of locations and at varying times to try to ensure that the variability within the population of interest is represented.

Generally, an intercept study entails selecting respondents by stopping them (i.e., intercepting them) in a public place (e.g., a shopping mall). Once a potential respondent is stopped by an interviewer, he or she can be screened for the criteria that determine the eligibility for the particular study. In a mall intercept study, qualified respondents are often taken to an interview facility within the mall to complete the main body of the questionnaire. Mall intercept studies are particularly useful in cases where respondents need to view or handle materials, and the method often can provide an easy and cost effective way to locate "live" respondents. Additionally, mall intercept studies can use longer and more complex questionnaires than could reasonably be implemented using a telephone or mail survey.

The key disadvantage of a mall intercept study is that it generally relies on nonprobability sampling for respondent selection, and therefore the results cannot be used to make statistical determinations about any identifiable target population. Because respondents must be present at the mall (or public place) and generally are approached in a nonrandom fashion (i.e., one that is at the convenience of the interviewer), the researcher cannot definitively determine how well the sampled respondents represent any population, not even the population of those at the mall on the day(s) of data collection. Research on mall intercept surveying has shown them to underrepresent lower-income and older persons. This does not mean that the results cannot be used to say *anything* about the population of mall attendees, only that the results cannot be used with confidence to make any statistically reliable estimates about them or any other population. For example, no statistical calculation of sampling error is justified when using a mall intercept sample, because there is no known nonzero probability of selection that can be assigned to the members of any target population.

A number of procedures can be put into place when using a mall intercept study that may help to increase the representativeness and reliability of the survey. First, it should be determined whether a representative proportion of the population of interest would actually be found at the mall or other selected intercept locations. For this reason, some intercept studies are not conducted at malls but are instead conducted outside of grocery stores, movie theaters, and/or other places where the target population is most likely to be found. Second, the survey researcher should carefully consider the geography and number of intercept locations at which data will be gathered. For example, to generate credible findings for a study of the adult population in the United States, it would not be wise to select only one mall in California and one in New York. Once the broad geographic areas are selected, the researcher ideally would randomly select the malls to be used for the study against bias being introduced in selecting only malls or locations of a particular quality or character. Third, interviewing should be conducted over a variety of days (weekdays and weekends) and should also be conducted at a variety of times of day and evening to ensure greater diversity in respondents. Fourth, a systematic sample of respondents at the mall(s) or other location(s) should be selected rather than allowing interviewers to approach respondents at the interviewers' own convenience. Finally, the researcher can consider the value of conducting a small-scale probability sample (using telephone or mail) to determine how the demographics of the mall intercept respondents and those selected at random differ, if at all, and whether or not these differences have any implications for the survey's findings.

Some mall intercept surveys also use quotas to ensure that the mix of survey respondents is more likely to represent the important subgroups in the population. These quotas are developed using external data (for example, census data or syndicated market research data) that indicate what the actual distribution of the total population looks like. When quotas are used, interviewers approach respondents who generally appear to fit the appropriate quota cell requirements (for example, 10 females and 8 males between the ages of 35 and 49) and may ask additional screening questions to determine whether the respondent does in fact fit the criteria.

In a mall intercept study, it is essential to use well-trained interviewers. The representativeness of the survey respondents can be increased if interviewers are trained to approach a variety of potential respondents, not simply those who look friendly or most approachable. Additionally, interviewers must be carefully trained to present visual or other materials to all respondents in a consistent fashion so that the stimuli shown to each respondent are comparable across interviews.

Although not appropriate in circumstances where it is necessary to provide precise statistical estimates with strong external validity (i.e., low chance of nonnegligible coverage error and/or nonresponse error), a well-designed and carefully conducted mall intercept study can provide useful results for the needs of many clients.

Sarah Butler

See also Convenience Sampling; Coverage Error; External Validity; Nonprobability Sample; Nonresponse Error; Probability of Selection; Probability Sample; Quota Sampling; Systematic Sampling

Further Readings

Hornik, J., & Ellis, S. (1988). Strategies to secure compliance for a mall intercept interview. *Public Opinion Quarterly*, *52*(4), 539–551.

Webster, C. (1996). Hispanic and Anglo interviewer and respondent ethnicity and gender: The impact on survey response quality. *Journal of Marketing Research*, *33*(1), 62–72.

Marginals

As it applies to survey research, a *marginal* is a number "at the margins" (at the edge or perimeter) of a cross-tabulation table of two or more variables. Statistical software that is used by survey researchers, such as SAS and the Statistical Package for the Social Sciences (SPSS), routinely create cross-tabs with the marginals showing as the default setting.

Table 1 shows a cross-tabulation between two variables—educational attainment (Not High School Grad, High School Grad, College Grad) and belief in the existence of extraterrestrial life (Believe, Not Sure, Do Not Believe)—from a survey conducted in 1996. This table displays the number of respondents (i.e., absolute frequency counts) that fall into each of the conditions (cells) shown in the table. The marginals in the Total column show the number of

Table 1 Cross-tabulation of education and belief in extraterrestrial life; *absolute* frequencies

	Not HS Grad	*HS Grad*	*College Grad*	*Total*
Believe	35	247	138	420
Not Sure	25	146	77	248
Do Not Believe	26	114	31	171
Total	86	507	246	839

Source: Buckeye State Poll, December 1996; Ohio State University Center for Survey Research.

respondents out of the total 839 surveyed who gave each answer when asked about whether or not they believed "in the existence of life somewhere else in the Universe than on Earth." Of the 839 adults surveyed, 420 said they did believe, 248 said they were not sure, and 171 said they did not believe. From the information in Table 1, one can quickly surmise that nearly half (420 out of 832) said they did believe. The marginals along the Total Row at the bottom of the table show the number of respondents among the total 839 in each of the three educational attainment categories. From these marginals, one can surmise that about 3 in 5 (507 out of 832) of those surveyed graduated from high school but not from college.

Table 2 shows the results of the same survey data except that it is the relative frequencies (percentages) that are displayed across the rows, down the columns, and as the marginals in each of the cells for the table. The top percentage in each cell is the row percentage for each educational level and the bottom percentage in each cell is the column percentage for each answer to the belief question. The margins, however, show the percentages for either that row or that column.

The information in Table 2 is more informative to a consumer of these data because it conveys information about both the answers to each variable and the interrelationship between the two variables in a way that is easy and quick to understand. For example, it is not clear from the results presented in Table 1 how the two variables are related, but from Table 2 it is readily apparent by looking at the row percentages that as educational attainment increases, so does the portion of people who believe in the existence of extraterrestrial life. Furthermore, as long as one knows the total number of respondents who were surveyed, then the absolute frequencies for each cell can be reproduced from the percentages in Table 2.

Table 2 Cross-tabulation of education and belief in extraterrestrial life; *relative* frequencies

	Not HS Grad (%)	*HS Grad (%)*	*College Grad (%)*	*Total (%)*
Believe	8.3	58.8	32.9	50.1
	40.7	48.7	56.1	
Not Sure	10.1	58.9	31.0	29.6
	29.1	28.8	31.3	
Do Not Believe	15.2	66.7	18.1	20.4
	30.2	22.5	12.6	
Total	10.3	60.4	29.3	100

Source: Buckeye State Poll, December 1996; Ohio State University Center for Survey Research.

Paul J. Lavrakas

See also Relative Frequency; SAS; Statistical Package for the Social Sciences (SPSS)

Margin of Error (MOE)

The margin of error (MOE) is a statistical concept that is most often associated with polls and pollsters. It serves to quantify the uncertainty associated with sampling in a poll or other survey. In survey research, it is almost never practical to measure the entire population. As a result, pollsters rely on random samples that are intended to be representative of the population. Because polls randomly sample from within a population, there will always be some amount of uncertainty, or variable error (variance), associated with their results. Simply put, if a U.S. pollster were to randomly sample 1,500 adults in a national survey, it is unlikely that these 1,500 people would perfectly reflect the opinions of the 200-plus million adults in the country.

The MOE can account only for random sampling error. It is unable to capture variance or bias that may be due to other aspects of total survey error, such as miscounts, incorrect coding, question bias, nonresponse caused by not gathering data from sampled respondents' when they could not be contacted or they refused to cooperate, and/or respondents lying or not answering all of the questions.

A real-life example illustrates the MOE's meaning and its use by pollsters and journalists. A Pew Research Center poll conducted October 27–30, 2004, asked respondents to identify for whom they were going to vote in the 2004 U.S. presidential election. The results found that 51% of respondents identified George W. Bush, 48% John Kerry, and 1% Ralph Nader. Pew reported that the sample size was 1,925 likely voters, with an MOE of approximately ± 2.5 percentage points.

The MOE is typically calculated based on one of three levels of confidence: 99%, 95%, or 90%. Pollsters most commonly rely on the 95% level of confidence. Roughly speaking, MOEs at the 95% confidence level are 24% smaller than at the 99% level if the sample sizes are the same (an MOE of approximately ± 1.9 at the 99% level of confidence would result in the example). When using a 95% confidence level, it is expected that the "true" percentage for the population will be within the MOE of the poll's reported percentage (i.e., the confidence interval) 95% of the time (19 times out of 20). Using the Pew poll example, this means that the true population's vote for Bush would have been expected to be between 53.5% and 48.5% (i.e., 51 ± 2.5), 95% of the time, had the same Pew survey been conducted many different times using different (but similarly designed) random samples of similar size.

In surveys that use a simple random sample, the MOE is easily calculated. At the 95% level, it is calculated by the following equation, $\pm 1.96(SQRT(PQ/(n))(100)$, where P represents the percentage of interest (e.g., 51% support for Bush in the 2004 Pew poll) and Q represents $1 - P$. The size of the sample on which the percentage is based is represented by n. The 1.96 is a constant associated with the 95% level of confidence. As the equation indicates, the MOE is very much affected by the survey's sample size. Thus, in the Pew example, had a simple random sample been used, the MOE would be calculated by $\pm 1.96(SQRT((.51)(1 - .51))/(1925))(100)$ or ± 2.2, which is slightly less than what Pew reported.

It rarely is possible to randomly sample from the entire population because pollsters usually do not have a complete list of everyone in that population. This typically results in the use of some sort of multistage sampling design, which ultimately affects the MOE, usually inflating it, as measured by the design effect (*deff*). Accordingly, there are a number of different ways to determine the MOE, all of which are dependent upon the particular sampling design used. Although these equations change, the essential importance of the sample size does not.

Often, the MOE is incorrectly interpreted. For instance, the results of the Pew poll did not indicate that Bush and Kerry were statistically tied. Nor did they indicate that Nader could have received 0% of the vote, even though the MOE on the Bush percentage was $\pm 2.5\%$; that is because there was a separate (and smaller) MOE on the Nader percentage. They also did not imply that a lead mattered only if it was greater than 2.5%. The MOE indicated only that the expected true percentage for the population of likely voters would be within the MOE of the poll's reported percentage X% of the time, where X represents the chosen confidence level.

James W. Stoutenborough

See also Confidence Interval; Confidence Level; Design Effect (*deff*); Multi-Stage Sample; *N*; Poll; Pollster; Population; Random Sampling; Respondent; Sample Size; Sampling Error; Simple Random Sample; Total Survey Error (TSE); Variance

Further Readings

Gawiser, S. R., & Witt, G. E. (1994). *A journalist's guide to public opinion polls*. Westport, CT: Praeger.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey Bass.

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Warren, K. F. (2002). *In defense of public opinion polling*. Boulder: Westview.

Mass Beliefs

The concept of "mass beliefs" refers to the norms, attitudes, and opinions held by the general public as opposed to those held by elites (e.g., politicians, journalists, and scholars). The term does not imply that all members of the public (masses) hold the same beliefs, but rather that certain beliefs are held in common by subsets of the general citizenry that are nonnegligible in size. Nowadays, surveys and polls often are the source of data that helps to define these beliefs and identify which subsets of the masses hold them.

Origins of the Concept of Mass Beliefs

The concept of "mass," articulated from the early through middle 20th century, refers to essentially rootless, alienated, and ignorant individuals, loosely affiliated in society and culture. Some theorists feared the masses' potential for mob violence, while others thought it necessary to foment political revolution among the masses. Generally, however, masses were thought to be docile and vulnerable to manipulation by political, economic, and military elite leadership and mass communication technologies. Elite images during the period ranged from positive to very negative.

The concept of "beliefs," from a psychological, cognitive perspective, refers to an individual's informational components about objects, persons, events, ideas, and actions used to understand his or her world. Beliefs often are assumptions, ranging from simple to complex, and may include evaluative dimensions. It is often thought that individuals prefer internally consistent beliefs in order to avoid tension. Beliefs do not exist in isolation but have personal psychological unity. Many basic beliefs are held in common by most individuals within a society and culture and usually are considered incontestable. However, other beliefs are not shared and are debated. Psychologists have explored beliefs in conjunction with values, attitudes, and actions.

Political psychology addresses linkages between beliefs and social action. For example, there is widespread U.S. support for the sociopolitical value of humanitarianism, which advocates that more fortunate persons, groups, and nongovernmental organizations ought to provide assistance to the needy. There is tension between the values of humanitarianism and egalitarianism: Egalitarianism advocates direct government intervention, whereas humanitarianism stresses voluntary assistance. Another example of tension that affects social action is illustrated in conflicts between values of social conformity, desire for social order, and the wish to maximize personal autonomy. These tensions, thought to generate perceived threat within the personality structure, are said to contribute to a mass belief impetus toward authoritarianism and limits placed upon civil liberties. The phenomenon partially explains prejudice and political intolerance against nonconforming groups. Evidently, a majority of individuals prefer social conformity and have strong impulses to enforce social cohesion via mandated obedience to political authority. Linkages between mass beliefs, values, and actions explain the impulse to punish nonconformists who challenge social norms. Individuals incorporate group-centrism into their thinking.

The Industrial Revolution of the late 18th and early 19th centuries disrupted traditional agrarian ways of life that revolved around small, isolated cultures. Greater societal dependence upon machine production of goods led to demand for unskilled labor. Subsequent dislocations were caused by large-scale migrations as people moved around the globe. Humans increasingly experienced class divisions, urbanization, and mass advertising. They consumed machine-produced goods. Growth of specialized industries and subsequent interdependencies led to losses in self-sufficiency. People experienced a quickened pace of life due to new transportation modes and communication technologies. Increasingly remote elites exercised monopolies in the organization and control of dominant forms of communication media and provided largely one-way message transmissions to the populace. Populations were eager to receive information and entertainment messages, even though the messages might be biased and/or debasing.

Critics decried what they characterized as totalitarian communication technologies that served ruling class desires and decisions. They characterized the mass audience as relatively passive, submissive, reactive, irrational, atomized, and helpless. The masses' desire for easy living and love of amusement was said to be among the minimal bonds that connected people. Politically, citizens were characterized as isolated and dispersed. Their input into a democratic system was conformist at best, their influence on major societal decisions was negligible, and their impact on issues, policies, regulations, and laws was illusory and mythical. Angst was expressed by reformers who concocted schemes to unite and mobilize the masses. Visionaries believed the masses shared an underlying but latent bond due to their interests, environment, deprivation, exploitation, and nationality that needed to be made salient before they could be mobilized into action.

Commentators with historical perspectives were at the other extreme. They argued that humans always have been susceptible to the lowest common denominator and to pressure from special interest groups, whether in culture, politics, war, or other social endeavors. Defenders of modernity optimistically saw new technologies as offering greater choice and richer intellectual variety at a cheaper cost to a greater proportion of the population than ever before, thus enabling humans to react to their unique situations and circumstances in individualized, meaningful, and less predictable ways. Accordingly, the optimists did not believe conspiracy theorists who argued, for example, that the mass communication media were ruled by a nameless, sinister elite cabal.

Various critics were concerned particularly with mass communication media effects, which led to a robust debate concerning the presumed linkage between "mass" and "beliefs." Their arguments often turned on whether or not a steady diet of propaganda and vulgar cultural messages, available to millions nearly simultaneously through relatively anonymous and inaccessible elite communicators, had led to assembly-line, undifferentiated individual thought and behavior.

Explanations for Mass Beliefs

Once the concept of mass beliefs was considered viable, it was of great interest to explain its origins. The answer often varied according to the theoretician's or researcher's academic discipline. Some political scientists fingered the political elite and the governance system as primary sources. Economists readily pointed to systems of production and consumption, supply and demand, or to cost/benefit principles. Various sociologists blamed the purpose and performance of mass education and the effectiveness of propaganda as causes of the public's inferior socialization vis-à-vis desirable social and political roles. Communication scholars sometimes worried that the mass communication media had inordinate influence. Theologians were sensitive to the impact of spiritual entities or the ministrations of religious institutions or cults. And so on. Perhaps the most fully articulated discussions of mass beliefs have come from psychology and political science.

In the 1960s, University of Michigan researchers posited that the concept of ideology would be helpful in summarizing and explaining the thousands of beliefs that an individual might hold. Similar to the social psychological concepts of attitude and value, ideology combined individual beliefs into coherent functional structures. Ideology's utility exists in its capacity to bundle beliefs in an organized and predictable manner, reducing the sheer number of beliefs to a manageable number of demonstrable principles that can dominate emotional responses, cognitive thinking, and action. However, most Americans were not believed to possess an ideology. Philip Converse said in the early 1960s that mass beliefs as expressed through ideology were largely unimportant in American politics and that numbers alone did not create political power. His view did not rule out ideology as a driving force. While ideology was not a mass characteristic, it was characteristic of political elites, who behaved accordingly, using central ideological principles and information that logically related to the exercise of power. Because ideology did not apply to everyone, the idea of "constraint" was advanced. *Constraint* referred to the extent to which an individual's beliefs were interrelated and the degree to which one belief could be predicted if another were known. Disputed research analyses from the 1970s, relying on data collected during turbulent times that stimulated new issues (the Vietnam War, Watergate, race conflicts, urban crises, economic instabilities), suggested that a newer generation of voters generally was more aroused. Contributory factors were said to include weakening party identification, dissatisfaction with political processes, increases in education that correlated with ideological perspectives, and increases in coherent and consistent issue voting.

Challenges to the hypothesis of minimal constraint in mass belief systems (unconnected beliefs, ideas, attitudes, and positions across issues) have been voiced since the 1980s. Those against the hypothesis argued that voters reason about parties, issues, and candidates and make inferences from observations gleaned from mass media, political party campaigns, and informal opinion leaders. Knowledge about the past and present and projections to the future play into voter decisions. Compared to elites, such reasoning probably is based on lower levels of information and political understanding, because voters use informational shortcuts to simplify thought processes. Recent theory downplays the electorate's mass nature and focuses on a more educated polity that pays attention to issues that are personally relevant and well thought out. Converse published a rejoinder in 2000, noting that the average level of electorate information was low, but there was high variance that could be explained by ability, motivation, and opportunity. He pointed out that degree of ideology and information diminished rapidly as one moved from the elite to the masses, except for lingering affective traces. Today, the conceptualization of voters as motivated by issue interest groups and ideological stances has led to more concern for how a diverse electorate is fragmented or compartmentalized (for example, along racial lines). Fewer theoreticians view society as divided dichotomously into power elite versus the masses.

Whether or not the population is passive or active has been a key consideration in the modern variant of the earlier debate. Widespread adoption of converging computer, Internet, and media technologies has given individuals and groups the potential freedom to interact on a global scale, freeing them from a strictly consumer role to a producer role that is equally (or more) attractive. In contemporary debate, Robert D. Putnam's "bowling alone" perspective sees socially isolated people as lacking social capital and sometimes uncritically accepting whatever is suggested, adopting the course of least resistance. Another voice articulates the vision of a somewhat antiauthoritarian "brave new virtual community" of Internet enthusiasts who dynamically, voluntarily, critically, and cooperatively search for and/or produce particularized, selected cognitive information and who relish a variety of emotional sensations. Whether one or the other critical perspective will be vindicated depends in large part on whether future communication technologies are controlled by elite, powerful, centralized, and top-down organizations and institutions or by a decentralized system that allows access so that ordinary users can build social capital, become more energized in social processes, and can exercise more individual and collective power.

Ronald E. Ostman

See also Attitudes; Political Knowledge; Public Opinion; Social Capital

Further Readings

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American voter*. New York: Wiley.

Carmines, E. G., & Stimson, J. A. (1982). Racial issues and the structure of mass belief systems. *Journal of Politics*, *44*(1), 2–20.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press of Glencoe.

Converse, P. E. (2000). Assessing the capacity of mass electorates. *Annual Review of Political Science*, *3*, 331–353.

Feldman, S., & Steenbergen, M. R. (2001). The humanitarian foundation of public support for social welfare. *American Journal of Political Science*, *45*(3), 658–677.

Glynn, C. J., Herbst, S., O'Keefe, G. J., & Shapiro, R. Y. (1999). *Public opinion*. Boulder, CO: Westview.

Lowery, S. A., & DeFleur, M. L. (1988). *Milestones in mass communication research: Media effects* (2nd ed.). New York: Longman.

Mills, C. W. (1956). *The power elite*. New York: Oxford University Press.

Nelson, T. E., & Kinder, D. R. (2001). Issue frames and group-centrism in American public opinion. *Journal of Politics*, *58*(4), 1055–1078.

Nie, N. H., Verba, S., & Petrocik, J. R. (1976). *The changing American voter*. Cambridge, MA: Harvard University Press.

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Rosenberg, B., & White, D. M. (Eds.). (1957). *Mass culture: The popular arts in America*. New York: Free Press.

Matched Number

A *matched* telephone number is one that has a mailing address associated with it. Typically, it also has a name matched to it. The majority of matched telephone numbers are also listed telephone numbers, but some are unlisted. Unlisted numbers (those not listed with directory assistance or published in any local telephone book) can be matched to an address (and possibly a name) because the commercial vendors that perform the matching use databases that contain some unlisted telephone numbers with addresses and names, such as those that can be retrieved from public records in many states (e.g., vehicle registration lists, public tax bills, and other public records and databases). However, this matching process is not 100% reliable, since people often move or change their telephone numbers.

Whether or not a telephone number can be matched is predictive of the likelihood that a completed interview will be attained with that household in a telephone survey. A greater proportion of interviews are completed with numbers that are matched than are completed with unmatched numbers. A primary reason for this is that matched numbers have an address associated with them. As such, researchers can send advance mailings to these households when they are sampled for a telephone survey to alert them ("warm them up") to the fact that an interviewer will be calling them. Advance letters with as small a cash incentive as $2 have been found to raise cooperation rates by approximately 10 percentage points in general population telephone surveys in the United States. Another important reason that cooperation rates in telephone surveys are higher for matched numbers is that those whose numbers are able to be matched are generally less likely to regard a telephone interviewer contacting them as an invasion of their privacy.

On average, matched telephone numbers require fewer callbacks than unmatched numbers to reach a proper final disposition. Thus, the calling rules used by a survey center to process matched numbers should differ from the rules used to process unmatched numbers. However, unless a survey center has its telephone samples screened for matched/unmatched status or receives this information for each number in the sample from its sample vendor, it will not be possible for the survey center to take the matched/unmatched status into account as their computer-assisted telephone interview (CATI) system processes the callback attempts.

Paul J. Lavrakas

See also Advance Contact; Advance Letter; Calling Rules; Cold Call; Computer-Assisted Telephone Interviewing (CATI); Listed Number; Random-Digit Dialing (RDD); Telephone Surveys

Further Readings

Camburn, D., Lavrakas, P. J., Battaglia, M. P., Massey, J. T., & Wright, R. A. (1996). Using advance respondent letters in random-digit-dialing telephone surveys. *Proceedings of the Section on Survey Research Methods* (pp. 969–974). Alexandria, VA: American Statistical Association.

Shuttles, C. D., & Lavrakas, P. J. (2004). *Two advance letter experiments to raise survey responses rates in a two-stage mixed mode survey*. Paper presented at the Joint Statistical Meetings, Toronto, Ontario, Canada.

Mean

The mean is a descriptive statistic that survey researchers commonly use to characterize the data from their studies. Along with the median and mode, the mean constitutes one of the measures of central tendency—a general term for a set of values or measurements located at or near the middle of the data set. The arithmetic mean is the most commonly used measure of central tendency and is what is commonly referred to as the "average" of the data values. The mean is calculated by taking the sum of the data set and dividing by the number of observations to obtain the arithmetic mean. For example, in a data set containing the values 1, 2, 3, 4, 5, 6, 7, 8, and 9, the arithmetic mean would be calculated by adding up the data values—45 in this instance—and dividing by the number of observations—9 in this instance. In this example, the arithmetic mean is equal to 5.

Since the mean takes into account all of the available data within the data set, the mean is highly influenced by outlying data points (outliers). Thus, the median is often used when a data set has outlying data points that could influence the mean and misrepresent the data set. However, it is possible for the mean and median to be equal, for example, in data sets in which the data are normally distributed. The mean is valid only for interval and ratio and not for ordinal and nominal data.

There are many other types of means that can be calculated, including geometric, weighted, harmonic, and so on. The choice of the most appropriate mean to use depends on the nature of the data available. For instance, a geometric mean is commonly used when the data are interpreted according to their product and not their sum. This would be useful when calculating the average rates of annual return in stock investments, when numbers are reported as multiples of the base number. However, these other types of means typically are not used in survey research as much as the arithmetic mean.

Richard Kwok

See also Interval Measure; Median; Mode; Nominal Measure; Ordinal Measure; Outliers; Ratio Measure

Further Readings

Kornegay, C. (1999). *Math dictionary with solutions: A math review*. Thousand Oaks, CA: Sage.

Mean Square Error (MSE)

The problem with speaking about the average error of a given statistical model is that it is difficult to determine how much of the error is due to the model and how much is due to randomness. The mean square error (MSE) provides a statistic that allows for researchers to make such claims. *MSE* simply refers to the mean of the squared difference between the predicted parameter and the observed parameter. Formally, this can be defined as

$$E[(\theta^* - \theta)^2]. \qquad (1)$$

In Equation (1), E represents the expected value of the squared difference between an estimate of an unknown parameter (θ^*) and the actual observed value (θ) of the parameter. In this instance, the expected value of the MSE simply refers to the average error one would expect given the parameter estimate. MSE is often categorized as a "loss function," meaning that it represents how wrong the estimated parameter actually is, allowing one to then calculate the parameter's impact on the rest of the model. However, unlike other loss functions, MSE is convex everywhere.

Substantively, the MSE value can be interpreted in many different ways. Statistically, the goal of any model should be to reduce the MSE, since a smaller MSE implies that there is relatively little difference between the estimated and observed parameters. Generally speaking, a well-fitted model should have a relatively low MSE value. The ideal form has an MSE of zero, since it indicates that there is no difference between the estimated and observed parameters. This means that a relatively low MSE should be somewhat close to zero. This interpretation can also be used to compare competing models, using the MSE value as a rubric for deciding which model is best. The model that has the lowest MSE should be considered to be the best, since it provides the best fit and provides the least biased estimate. However, MSE should be used in conjunction with other statistics, such as Adjusted-R^2, in order to ensure that the researcher is choosing the best possible model.

MSE is also valuable when it is thought of as a composite of the variance of the estimated parameter and some unknown random bias. Specifically, this can be defined as

$$\mathrm{Var}(\theta^*) + (\mathrm{Bias}(\theta^*, \theta))^2. \qquad (2)$$

Using Equation (2), we can say that an unbiased parameter estimate should have an MSE equal to the variance of the estimated parameter, whereas a biased parameter estimate will have a residual value that represents the squared parameter bias. This is helpful in terms of model building since it allows the researcher to speak in terms of the variance explained by the model and the variance left to random error. A model that has a nonzero bias term can be somewhat problematic since the MSE value serves as the basis for the coefficient standard error, which is then compared to the coefficient magnitude to create the t statistic. A biased MSE can affect these estimates in many ways.

A positive bias term implies that the estimated value is higher than the true value ultimately drawing the t statistic closer to zero, resulting in an increase in Type II error. A negative bias term implies that the estimated value is lower than the true value, which pushes the t statistic away from zero, resulting in an increase in Type I error. Additionally, a relatively low MSE value does not necessarily imply that the parameter estimate is unbiased, since a relatively high bias term can be compensated for by a minimal variance in the estimated parameter. All of these things should be kept in mind when using the MSE value for variable selection and model comparison.

However, when determining how well a statistical model fits the data, MSE can be a valuable tool, because it allows one to calculate the average error that the parameter estimate produces, which can then be partitioned into the variance of the estimated parameter and some bias term. With MSE, one can compare one's present model error to the error that one would expect given the data, which is useful for interpreting a model's explanatory power as well as comparing it to other models that attempt to achieve the same end. Ultimately, MSE can be used to help minimize the errors of a given model and is one of many tools that survey researchers and other social scientists use to conduct meaningful quantitative research.

Bryce J. Dietrich

See also Confidence Interval; Random Error; Significance Level; Standard Error; Standard Error of the Mean; *t*-Test; Type I Error; Type II Error; Unbiased Statistic; Variance

Further Readings

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13, 469–475.

Greene, W. (1999). *Econometric analysis* (4th ed.). New York: Prentice Hall.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge: MIT Press.

Measurement Error

Measurement is the assignment of symbols, usually numbers, to objects according to a rule. Measurement involves both creating a rule and making assignments. The symbols to be assigned represent attributes of the object. *Error* in measurement is any deviation of the assigned symbol from the "true" value that should be designated to the object. A term that is used to refer to how accurately something is measured is *construct validity*.

For example, a researcher might want to measure a person's level of education. In this case, the person is the "object" and level of education is the "attribute" for which the researcher wants a value assigned to each object. The goal of measurement is to assign to the person a symbol—a number—that represents her or his true educational attainment. In order to achieve this goal, the researcher needs first to define education and its range of values. Then the researcher needs to devise a method to designate a value of education for the person. There are myriad ways to do this, including observing the person's dress or behavior, documenting the vocabulary the person uses in everyday discourse, retrieving information from school records, testing the person's knowledge of various subjects, or asking the person to report how many years of schooling she or he completed. The information obtained then is converted to a value or category of education.

Understanding the Meaning of Error

Errors can be made at any or all of these steps. Education may be defined incorrectly, either at the conceptual or the operational level. An incorrect conceptual definition of education is a theoretical mistake—a misunderstanding of what education means in the context of theory construction. For example, education might be conceived by a researcher as a person's "mastery of correct social appearances." An incorrect operational definition is an error in the way education is conceived for the purpose of gathering information. An operational definition corresponding to the "social appearance" conceptual definition might be "the extent to which a person exhibits behaviors that are seen as refined in a given culture." This operational definition would lead to seeking information about people that would allow them to be placed in categories of social sophistication.

Errors in conceptual definitions are identified and debated in theoretical discussions of the concepts (constructs) that are employed in social research. A researcher who argues that *education* means "a mastery of social appearances" may be confronted by another who claims that education really concerns "the accumulation of knowledge." The research community debates the most useful conceptual definitions of the construct. Such debates take place as research, employing alternative conceptual definitions of key concepts, is carried out. Theoretical arguments and empirical data lead the research community to adopt some conceptual definitions as correct and to treat others as erroneous. Measurement begins with reasoned decisions about the essence of the concept to be gauged.

This theoretical beginning is essential to the measurement process, but discussions of measurement and measurement error typically start at the operational definition phase. The conceptual definition is taken as established. Attention is focused on how well the theoretical idea is translated into more concrete language that allows information about people to be gathered and used to assign them to categories. The issue of measurement error becomes, "How well do the adopted operational definition and its attendant procedures capture the concept of interest?" There may be a mismatch between the concept and its operational definition. Education, conceptually defined as "the accumulation of knowledge," might be operationally defined as "performance on an examination of arithmetic and reading skill." The education scores resulting from the examination may be criticized because the test does not address other significant areas of knowledge and learning. To the degree that the operational definition fails to capture the theoretical concept, the resulting measure is termed *invalid*.

But the validity of a measure depends not only on the operational definition. It also depends on the methods employed to gather information and the way that information is used to categorize people. In general, a measure is said to be *valid*—to have strong construct validity—if it measures what it claims to measure. An operational definition may translate the abstract concept adequately so that observations can be made, but the observations themselves may be poorly conceived.

For example, a test of knowledge to measure educational attainment may cover all significant areas of knowledge, but the individual questions in the test instrument may be confusing. Or, the questions may be appropriate, but the way in which scores are combined in categories may give undue weight to some sorts of knowledge. Any of these sources of error—mismatch between the concept and the operational definition, ill-formed observational techniques, or misguided scoring methods—can affect the validity of a measure. For a measure to be truly valid, it must be error-free. In practice, however, the best researchers can claim is that one measure is more valid than others. No measure is completely error-free.

The types of error affecting a measure's validity can be divided into two classes—*systematic* error (or bias) and *random* error (or variance). In order to identify these types of error, researchers need to be able to observe how a measure performs in repeated trials with the same people in the same conditions. Systematic error is the tendency for a measure to produce scores that are consistently different from the true score in one way or another. A measure is biased if, in repeated applications, its scores tend to deviate from the true score in one direction. This might occur, for example, if a particular measure of education tended to systematically over- or underestimate the level of attainment. Another form of systematic error occurs when the scores produced by a measuring technique have errors that are correlated with the true value of the variable. This might occur if a measure of education tended to consistently underestimate the values for people with higher educational attainment and to consistently overestimate the values for people with lower levels.

A measure is affected by random error if, in repeated trials, its scores deviate (vary) from the true score with no consistent pattern. The results of repeated applications produce scores that are "all over the map," not clustering at one level or another. The less a measure is subject to random error, the more *reliable* it is said to be. Reliable measures are ones that tend to produce consistent scores over repeated trials, even if those scores are not actually valid ones.

Summarizing terms to this point, a measure is valid to the degree that it is error-free. Two types of error are threats to validity: systematic and random. Systematic error, bias or correlated error, decreases the validity of a measure because the scores produced are consistently wrong. Random error decreases validity because the scores produced are inconsistently wrong. Measures that are consistent are called "reliable." For a measure to be valid, it must be consistent and free of bias. It is possible for a measure to be reliable but not valid—if it is consistent but also has systematic error. It is possible for a measure to be free of systematic error but not valid—if it is not reliable. And, of course, it is possible—and all too likely—for a measure to be both unreliable (inconsistent) and contaminated by systematic error.

In discussing measurement error, one is concerned with the validity of survey measures. In order to assess validity, one needs to have some idea of what the "truth" is. Two ways of thinking about truth are common in survey research. First, survey measures are often compared to more tangible nonsurvey evidence to assess their validity. For example, survey reports of voting or hospitalization or victimization or expenditures or employment can be matched against records (so-called record checks) or observations of these states and behaviors. The information contained in the records or gleaned from observations is treated as true. If survey reports differ from the record or observational information, the survey measure is regarded as invalid. This way of thinking about truth is sometimes called *Platonic* (a reference to the Greek philosopher, Plato); it assumes that there is a real, objective, factual, true state of things in the world. The validity of a measure depends on its ability to capture this reality. For the variety of surveys that seek to measure behavior or states that can be observed outside the survey context, the Platonic conception of truth is common in discussions of measurement quality.

The second way of thinking about truth in survey research is sometimes called *psychometric*. This conception of truth dominates in discussions of the quality of measures of mental constructs that cannot be observed. Surveys often ask people to report their beliefs or feelings or perceptions or attitudes. There are no records that can be checked to verify self-reports of these cognitive matters, nor can they be observed objectively. How then to think about "the Truth" in this context? Psychometric true scores are defined in statistical terms as the expected value of a hypothetical, infinite set of reports by a "fixed" person. Imagine that a person could report his or her attitude by selecting a category from a list of options over a multitude of times, with no memory of previous reports, in circumstances in which the attitude does not change. The multitude of responses would vary to some degree, with some categories being selected more frequently than others. Among the categories selected frequently, one would be the most typical response, or the *expected value*. This category is the person's true score. Responses that deviate from this category are less valid—have more measurement error—to the degree that they deviate. Thus, there are two ideas of truth, one often applied to measures of phenomena that can (at least in theory) be observed without reliance on self-report, and one applied to phenomena that only can be measured through self-report.

As noted earlier, the validity of a measure depends on its operational definition, the methods of collecting information from people that are employed, and the method of assigning scores to them. Operational definitions may lead to collecting correct or incorrect sorts of information. Measurement error can arise from a number of sources in surveys, including the ways in which questions are constructed, the ways in which they are presented to respondents, and the ways in which respondents deal with them.

Survey Data and Measurement Error

Surveys consist of questions posed to people in order to elicit responses. Questions require respondents to comprehend and interpret their meaning, to retrieve relevant information, and to formulate and communicate a response. The way in which questions are written affects the ability of respondents to perform these tasks adequately. Measurement error can result when the respondents (a) do not understand what is being asked, (b) when they fail to retrieve relevant information from which to construct an answer, or (c) when they do not communicate an accurate response.

Vocabulary is an essential consideration when researchers are formulating the items to use in their questionnaires. Questions that employ obscure or difficult terminology are more likely to be misunderstood. Researchers too often erroneously assume that people are as familiar with the particulars of a survey topic as they are. This assumption can lead to the use of question terms that are unfamiliar or ambiguous to respondents. To the degree that this is so, respondents will have trouble comprehending what is being asked. Comprehension difficulties lead to erroneous responses.

Syntax also is an important question characteristic. Questionnaire items that are constructed with multiple clauses, with contingent conditions, may be confusing or may require a greater degree of attention from respondents. If respondents need to sort through possible meanings of what is being asked, they will need to expend more effort. If they fail to do the extra work, erroneous responses can result.

Task difficulty (i.e., respondent burden) must be considered when constructing questions. Independent of the words used or the syntax of the question, the kind of information requested will affect how well respondents perform. Some questions ask respondents to recall events in their lives. The further removed in time these events are, the less likely respondents will be to recall them or to report them within the correct time frame. This is particularly true of events that are not salient. For example, asking a respondent what he or she purchased at the grocery store a month ago presents a task that is very difficult if not impossible to perform accurately. Asking an adult to report on events that occurred in childhood is another example of a task that is fraught with difficulty.

Apart from burdens on memory, there are other sorts of task difficulty that can be posed by questions. For example, some survey items ask respondents to perform quasi- or actual arithmetical calculations—for example, to report what percentage of their time they spend on different activities, or to report their net worth. To answer such questions with any degree of accuracy would require substantial time and calculation. A single survey question does not provide the requisite conditions for an accurate response.

Another form of difficulty is posed by questions that ask respondents to report on behaviors or feelings that are socially disapproved. It is harder to admit to being lazy or to harboring prejudicial feelings than it is to report on completed projects and good will toward others, especially when a questionnaire is administered by an interviewer. Generally speaking, saying "Yes" (i.e., acquiescing) may be easier than saying "No" for some people. Norms of social desirability may also vary by culture, so task difficulty needs to be anticipated with this factor in mind.

Questions also present respondents with different kinds of response alternatives or options that may lead to measurement error. Open-ended questions require respondents not only to retrieve and organize relevant information but also to express themselves in ways that they think are responsive to the queries. It is frequently argued that respondents should be allowed to "speak for themselves" in this way rather than being confined to selecting a response from a list of categories. But the freedom to formulate a response to an open-ended question can involve considerable cognitive effort on the part of a respondent, particularly if the question concerns a topic that is not familiar. Articulation of an open-ended response will also be more or less difficult for people who are taciturn or gregarious in general.

Closed-ended questions ask respondents to choose a response from a list of pre-set categories. In general, this task may be an easier one for respondents than formulating an open-ended response, because it is clearer what sort of answer is expected. On the other hand, respondents can be influenced in their response selection by the range of alternatives offered. Additionally, the list of options may be discordant with the body, or stem, of the question. For example, a question may ask a respondent how frequently he or she has what is, in general, a very rare experience. Respondents may also find themselves without a category that appears to correctly map onto their experience or feelings. Some respondents may believe their response best fits between the categories that are offered to choose from. Finally, offering or withholding a "Don't know" option for selection among the list of response alternatives can have a profound effect on the number of respondents who will provide this response.

Individual questions are set among others in a questionnaire. The context within which a given question appears can affect responses to it. For example, if respondents are asked for their attitude toward a particular government policy in a series of questions, their responses to later items may assume that the earlier answers are taken into account. Another sort of context effect can occur when questions placed later in a long questionnaire do not receive the same level of attention that the lead-off items got. Fatigue may lead respondents to minimize effort in the later questions.

These are some of the ways in which question wording, form, and context can lead to measurement error. The way in which questions are communicated (i.e., the survey mode of data collection) to respondents can also have an effect on measurement validity. Respondents can encounter questions in a face-to-face interview, a telephone contact (either via an interviewer or Interactive Voice Response technology), a paper-and-pencil form, a laptop computer instrument, or an Internet survey. When respondents speak directly to interviewers, their answers may be affected by what they think the interviewer expects or will approve. When respondents complete a questionnaire without interviewer involvement, their responses can be shaped by the layout of and graphics used in the self-administered form. Both interviewers and respondents can make errors in the way that they record responses—for example, failing to record responses verbatim or ignoring a questionnaire skip pattern.

Measurement error can also occur when survey responses are processed at the end of a study. Responses to open questions need to be summarized in categories. The coding process can misplace a respondent's intended answer or it may not find a place for it at all. Achieving reliable placement of similar responses across coders can be an arduous and expensive process. Closed questions have many fewer coding problems, but they are not immune to data entry errors.

A substantial part of the field of survey methodology is about identifying, reducing, and/or correcting measurement error. Given the complexity of survey investigation, involving the abstract realm of conceptualization and the intricate practicalities of getting information from respondents, confronting measurement error is an enormous task. But while measurement error will never be eliminated, it is becoming better understood. And the study of the phenomenon itself provides much insight into human behavior. Understanding measurement errors teaches us truths about how people think and communicate.

Peter V. Miller

See also Acquiescence Response Bias; Bias; Closed-Ended Question; Coder Variance; Coding; Cognitive Aspects of Survey Methodology (CASM); Construct Validity; Don't Knows (DKs); Forced Choice; Gestalt Psychology; Graphical Language; Interviewer-Related Error; Mode of Data Collection; Mode-Related Error; Open-Ended Question; Primacy Effects; Questionnaire Design; Questionnaire Length; Questionnaire-Related Error; Question Order Effects; Random Error; Recency Effects; Record Check; Respondent Burden; Respondent Fatigue; Respondent-Related Error; Response Alternatives; Self-Reported Measure; Social Desirability; Systematic Error; Total Survey Error (TSE); True Value; Validity; Variance

Further Readings

Alwin, D. (2007). *The margins of error: A study of reliability in survey measurement*. New York: Wiley.

Biemer, P. P., Groves, R. M., Lyberg, L., Mathiowetz, N., & Sudman, S. (2004). *Measurement errors in surveys*. New York: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
Lyberg, L. E., Biemer, P. P., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N., et al. (1997). *Survey measurement and process quality*. New York: Wiley.

MEDIAN

Median is a descriptive statistic that researchers commonly use to characterize the data from their studies. Along with the mean (average) and mode, the median constitutes one of the measures of central tendency—a general term for a set of values or measurements located at or near the "middle" of the data set. The median is calculated by sorting the data set from the lowest to highest value and taking the numeric value occurring in the middle of the set of observations. For example, in a data set containing the values 1, 2, 3, 4, 5, 6, 7, 8, and 9, the median would be the value 5 as it is the value within the data set that appears in the middle—with four observations less than and four observations greater than the median value. The median can also be thought of as the 50th percentile.

It is possible that a data set can have a median that is not a specific observation within the data set. This happens when the data set has an even number of observations. In this instance, the median would be the mean of the two middle numbers. For example, in a data set containing the values 1, 2, 3, 4, 5, and 6, the median would fall between the values 3 and 4. In this instance, the median would be 3.5. There are three observations less than and three observations greater than the median value.

Unlike the mean, the median is not influenced by extreme outlying data points within the data set. For instance, in a response to a survey question about annual personal income, if one respondent reports an income that is 10 times greater than the next closest person, this respondent would be an outlier and would skew the mean value upward. However, the median would be unaffected by this outlier and would more accurately represent the middle of the data set. Thus, the median is often used when a data set has outlying data points that could influence the mean and thereby misrepresent the middle of the data set. This also is common in survey questions on home prices or issues related to costs and finances, when extreme outliers can dramatically affect the mean value. In this instance, presenting the median value would be much more informative about the average value of housing than the mean, as the median is not influenced by the outlying values.

Richard Kwok

See also Mean; Mode; Outliers; Percentile

Further Readings

Kornegay, C. (1999). *Math dictionary with solutions: A math review*. Thousand Oaks, CA: Sage.

MEDIA POLLS

News organizations conduct or sponsor public opinion research as part of their ongoing news coverage, including but not limited to election campaign coverage. Media polls, also called "news polls," have attained wide prevalence as a reporting tool but also remain a focus of debate and occasional controversy. Media polls are a central part of what Philip Meyer has explained as "precision journalism."

News may be defined as timely information, professionally gathered and presented, about events and conditions that affect or interest an audience. It can include what people do and also what people think. Polls provide a systematic means of evaluating both, elevating anecdote about behavioral and attitudinal trends into empirically based analysis.

News organizations long have reported—and continue to report—characterizations of public preferences without the use of polls, relying on expert (and sometimes inexpert) evaluations, informed (and sometimes uninformed) speculation, punditry, proselytizing, and conventional wisdom. Rigorous polling improves on these.

The media also have reported polls (rigorous and otherwise), whether provided by outside sources (e.g., government, academic, or corporate entities, interest groups, and public relations firms); syndicated or circulated by independent polling companies (in some cases for promotional purposes); or self-initiated. Only polls in the last category are classified as media polls, as opposed more broadly to polls reported by the media.

Media polls may best be understood as a means of covering a news beat—the beat of public opinion. In a process that in many ways closely reflects other news reporting, media pollsters, in their reportorial role, go to their best sources, ask their best questions, take down the answers, and report what they have learned. A key difference from other reporting is in the selection of sources, which, rather than either event dictated or arbitrary, is based (in rigorous polling) on a random probability sample of the population under study.

Good-quality media polls, then, represent good news reporting. Different from the retrospective approach more common in academic research, they most often provide for the immediate and timely evaluation of current events, adding a unique perspective that actively informs the public discourse as it helps make sense of a central element of the world around us—public attitudes and behavior.

Media polls are best known for their use in political coverage, whether pre-election polls measuring support for candidates and the attitudes, impressions, and policy preferences that inform those choices; or, outside election cycles, ongoing measurements of views on political performance and policies. The former is epitomized by, but by no means limited to, so-called horse race measurements of opposing candidates; the latter, by presidential approval and related measures. Both are staples of political coverage.

News organizations, however, also conduct surveys across many other issues, measuring a range of experiences, circumstances, preferences, and behaviors. Venturing far beyond politics and policy, media polls are conducted on social and economic matters, including consumer confidence, the environment, lifestyles, health, sports, popular culture, religious belief and practice, race relations, entertainment, interpersonal relationships, and more. Some news organizations also undertake international polling, even in conflict areas where measurement of public attitudes can be difficult yet is particularly vital, such as Afghanistan and Iraq.

Media polls on elections and politics have been a particular focus of popular and academic debate, given the prominence of such surveys, the passions aroused by partisan politics, and the interest of political scientists in these matters. One line of argumentation suggests that such polls may influence the very attitudes they seek to measure and report, for example, through a supposed bandwagon or snowball effect, in which a candidate's reported popularity begets further popularity; or a suppression effect, in which eligible voters may be demotivated by pre-election polls reporting that their candidate is set to lose, or win, by a wide enough margin that their participation is moot.

Neither has a strong empirical foundation; candidate leads in fact change hands, and turnout rises and falls from jurisdiction to jurisdiction independently of the reported standings in any one contest on the ballot. Nor does denying the public information that is readily available to the campaigns and related interest groups seem preferable. (The media's tendency to focus on the horse race in covering election polls, to the exclusion of richer evaluative data, is a more persuasive concern.)

Broader criticisms suggest that media polls may "manufacture" opinions by measuring attitudes that in fact are nonexistent, lightly held, or ill founded; that they selectively accentuate issues through their choice of topics; or that they misreport attitudes through inadequate sampling methodology, ill-constructed questions, or ill-conceived analysis. None of these is specific to media polls per se, but to all polls; and all, again, can be answered. The absence of opinions on given questions can be expressed and tabulated; measurements of strength of sentiment can be included; choice of topics to cover is a feature of all news reportage, not news polls alone; and the quality of output is an individual- or product-level matter rather than an indictment of the enterprise overall.

However, media polls, like all polls, require careful scrutiny of their methodological and analytical bona fides. Indeed, given the credibility they lend data, news organizations that conduct or sponsor polls have a special responsibility to uphold the highest possible standards of methodological rigor. Some acquit themselves well. Others fall short.

A related aspect of media involvement in polls is in standards and vetting operations, in which news organizations set basic disclosure requirements and methodological standards for the survey research they will report, then undertake concerted efforts to ensure that any data under consideration meet those standards. This is a vital function, too long and still too frequently avoided, to ensure the integrity of news reports that incorporate polls and other data.

Ultimately, perhaps the best rationale for media polls stems from the fundamental premise of an independent news media. Non-news organizations often conduct polls in their attempts to influence the public discourse, and they will continue to do so—regardless of whether or not news organizations conduct their own polls. Political campaigns will measure their candidates' standings and the attitudes behind those preferences and use those data to direct and sharpen their messages and strategy. Interest groups, corporations, publicists, and others likewise will propose to the media that their data, as they produce and analyze it, merit reporting as news. And pundits will make their way into the news pages and onto the airwaves, pronouncing their views often without the benefit of empirical data of any sort.

Media polling interrupts this spin cycle. By conducting their own rigorous and responsible measurements of public attitudes and behavior—for example, in helping to explain an election "mandate" by using exit poll data—media organizations can fulfill an essential aspect of their responsibility to report the news independently and accurately.

Gary Langer

See also Approval Ratings; Bandwagon and Underdog Effects; Election Polls; Exit Polls; Horse Race Journalism; National Election Pool (NEP); Precision Journalism; Question Wording as Discourse Indicators

Further Readings

Gollin, A. E. (1980). Exploring the liaison between polling and the press. Polls and the news media: A symposium. *Public Opinion Quarterly*, *44*(4), 445–461.

Herbst, S. (1993). *Numbered voices: How public opinion has shaped American politics*. Chicago: University of Chicago Press.

Jacobs, L. R., & Shapiro, R. Y. (Eds.). (2005). Special Issue: Polling politics, media and election campaigns. *Public Opinion Quarterly*, *69*(5).

Kagay, M. R. (1991). The use of public opinion polls by *The New York Times*: Some examples from the 1988 presidential election campaign. In P. J. Lavrakas & J. K. Holley (Eds.), *Polling and presidential election coverage* (pp. 19–56). Newbury Park, CA: Sage.

Lavrakas, P. J., & Traugott, M. W. (Eds.). (2000). *Election polls, the news media and democracy*. New York: Seven Bridges/Chatham House.

Lavrakas, P. J., Traugott, M. W., & Miller, P. V. (Eds.). (1995). *Presidential polls and the news media*. Boulder, CO: Westview.

Meyer, P. (2002). *Precision journalism: A reporter's introduction to social science methods* (4th ed.). Lanham, MD: Rowman & Littlefield.

METADATA

There is no single definition that adequately describes metadata, though it often is referred to as "data about data" or is a description of data. In other words, *metadata* is a set of highly structured and/or encoded data that describes a large set of data. It explains the data to be collected, processed, and published and answers questions regarding every facet of the documented data. The data can be an individual data item or a collection of data items, with a primary purpose of managing, understanding, and facilitating data. Most important, metadata describes diverse data products by emphasizing the similarities between them, thus allowing people to understand the diverse data a certain organization has produced.

Metadata in Survey Research

In survey research, metadata plays a vital role; it is used to describe statistical data from survey conceptualization to data dissemination. Metadata can be very broad and encompasses populations, methods, survey instruments, analysis approach, results of instruments, and so on. Furthermore, it describes the information about the data, including variances, response rates, and response categories. Most survey researchers perceive this world as hierarchical; thus, a good data model must be able to handle data in a straightforward fashion as well as being able to describe data hierarchically. As a result, metadata should provide detailed information regarding variables while reflecting the data model structure.

Part of the survey process centers on raw data that are collected from respondents and converted into data sets. The entire conversion process needs data management. In general, good software packages need to deal with the data and metadata involved in the survey. Such software should be able to translate any data and metadata into any format when dealing with the survey process. Normally, data managers create metadata, and since metadata is expensive to create, a great need exists for understanding how the data would be used prior to its actual creation.

In addition, data and metadata producers both need to possess adequate communication channels between them. Metadata producers must follow certain standards. First, they need to thoroughly understand the data and be able to encode the data information. Usually, a single disk file is created for each metadata record, which describes one data set only. Second, the metadata file is arranged properly. Third, there is a need to verify the data and make sure it is adequately described.

Metadata also varies depending upon the type and context of data. In order to understand the data more completely, the following need to be addressed:

1. *A definition of survey variables*. Each variable must contain a valid value, a code, an identifying name, and so on.
2. *Data model*. This model describes relationships between the variables regarding their categories and hierarchies.
3. *Route instructions*. This defines order, such as how questions are asked.
4. *Relationships*. Whenever relationships impose restrictions on the values of variables, the restrictions must check the consistency among the collected data.

Cary Stacy Smith and Li-Ching Hung

Further Readings

Backstrom, C. H., & Hursh-Cesar, G. (1981). *Survey research* (2nd ed.). New York: Wiley.

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N., et al. (1997). *Survey measurement and process quality*. New York: Wiley.

Mohler, P., Pennell, B.-E., & Hubbard, F. (2008). Survey documentation: Towards professional knowledge management in sample surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (Chap. 21). New York: Psychology Press/Taylor & Francis Group.

Methods Box

A methods box is a short news story or sidebar that accompanies poll stories and provides methodological details and clarifications about the survey, including how the respondents were sampled, how the interviews were conducted, the process of weighting the results, and the survey's possible error.

Many newspapers include a boxed feature (sometimes in a smaller type size) alongside a major poll story, and most television networks include the equivalent on their Web sites with details about their survey methodology. For example, it is policy at *The New York Times* that a methods box accompanies all articles that focus on a *New York Times*/CBS News Poll or a *New York Times* Poll.

There are different rules at other papers. *The Washington Post* and the *Los Angeles Times* include the basic facts of the survey in the main story with some additional information at the bottom of the accompanying graphic. But an extensive and detailed description of the poll methodology appears on the papers' Web sites.

There are two main criteria in creating a method box. First, it needs to be written in plain English that any layperson can understand. Second, it needs to be statistically correct. It should satisfy the editors that it is literate and the statisticians that it is correct technically. It is not always easy to do both.

Different organizations offer varying levels of detail regarding their methodology, but in general, they adhere to the standards for minimum disclosure that is part of the American Association for Public Opinion Research's code of professional ethics and practices, available on its Web site. The National Council on Public Polls has put together a pamphlet, "Twenty Questions a Journalist Should Ask About Poll Results," which can be found on its Web site.

Standard elements of methods boxes include interviewing dates, the number and type of respondents, the mode of interviewing, sampling and weighting particulars, and the margin of sampling error. In addition to the method of interviewing, the methodology should also include a list of languages in which data were collected and the name of the company that may have conducted the fieldwork.

Some method boxes describe their margin of sampling error as based on a 95% confidence level. Because "95% confidence level" may not be a concept familiar to the public as a whole, an alternative procedure would be to explain what that actually means. For example, for a survey of approximately 1,000 respondents, *The New York Times*' methods box reads, "In theory, in 19 cases out of 20, overall results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults."

If the story involves frequent references to key subgroups, the margin of sampling error for those subgroups should also be included. For example, "For Democrats, the margin of sampling error is plus or minus 4 points and for Republicans it is plus or minus 5 points." In a pre-election poll, the subsample of "likely voters," if any, needs to be explained.

Methodological statements also explain that the margin of sampling error is only one possible source of error. Question wording and order are among the possible additional sources of error, along with nonresponse and coverage errors. Simply to characterize the sample as "RDD" could be insufficient for many laypeople, and so it could be more effective to describe what an RDD sample is—random-digit dialing, or telephone numbers randomly generated by computer using a complete or random list of residential exchanges, providing access to both listed and unlisted telephone numbers. Variables that are used to weight the final results should be identified.

Methods boxes published in the newspaper generally provide the reader with the Web site address for the complete results or list other resources for obtaining the full question wording and complete results. Identification of the director of the poll or any outside consultants who may have assisted in the project is often included in the methods box.

When poll stories are picked up from the newswire, the separate methods box is not always included, so basic details about the survey should also be in the body of a poll story. At minimum, the story should have the interviewing dates, the number and type of respondents, mode of interviewing, and the margin of sampling error.

Marjorie Connelly

See also American Association for Public Opinion Research (AAPOR); Confidence Level; Disclosure; Margin of Error (MOE); Media Polls; National Council on Public Polls (NCPP); Poll; Survey Ethics

Further Readings

American Association for Public Opinion Research: http://www.aapor.org

How the statewide poll was conducted. (2006, September 28). *The New York Times*. Retrieved April 14, 2008, from http://www.nytimes.com/2006/09/28/nyregion/29methodcnd.html?_r=1&oref=slogi

National Council on Public Polls: http://www.ncpp.org

Minimal Risk

Minimal risk is a concept that relates to the protection of human subjects and thus to survey ethics. In Chapter 45, Subpart A, Section 46.102, Paragraph (i) of the *Code of Federal Regulations*, the U. S. Department of Health and Human Services Office of Human Research Protections defines *minimal risk* as a category of research activities in which "the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those encountered in daily life or during the performance of routine physical or psychological examinations or tests." In general, a person who participates in a survey as a respondent is thought to be exposed to minimal risk. As such, minimal risk is a classification that is associated with nearly all scientific survey research studies, although many nonsurvey research programs also may be appropriately classified as having minimal risk.

Minimal risk status is not determined by a clearly objective standard that is subject to a claim made by a research investigator. Instead, formal classification of research as minimal risk is a judgment based on the assessment of an institutional review board (IRB), a group of at least five scientists and laypeople responsible for overseeing research activities at any research institution funded by the federal government or a federal agency and subject to the *Code of Federal Regulations* governing research. There are widely accepted guidelines that have been developed by various government-sponsored expert panels. For example, in 2001, an analysis conducted by the National Bioethics Advisory Commission encouraged IRBs to act blindly with respect to specific characteristics of research subjects (e.g., those having cancer) in interpreting or assessing a study's risk of adverse consequence to research participants and, instead, to be guided in assessment of research risk in terms of "common risks . . . for example, driving to work, crossing the street, getting a blood test, or answering questions over the telephone." At the same time, it is widely appreciated that an IRB should consider the severity of the harm should a potential adverse consequence from research participation occur. Similar review groups are used by many private sector survey organizations that have no federal funding.

In general, for the purposes of considering whether to classify a research program as minimal risk, it is useful to consider both the probabilities of adverse consequences and the severities or magnitudes of those consequences should they occur. This kind of decision calculus is depicted in Figure 1.

Figure 1 reflects the subjective nature of the assessment of minimal risk. It is within the purview of each

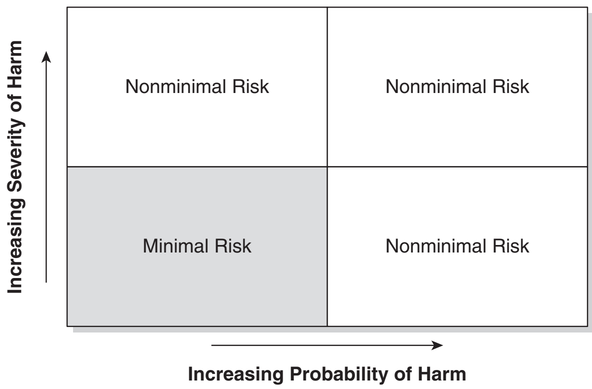

Figure 1 Probability and severity of adverse occurrences in determining risk level

individual IRB to determine the thresholds for likelihood of harm and severity of harm (should harm occur) that moves a research program beyond the realm of minimal risk.

It is generally understood that, for several reasons, a system for classifying research as minimal risk is a useful and important function of an IRB. These reasons include the following:

1. It helps the IRB be more efficient and effective in its oversight of research activities across its organization. Research programs not designated as minimal risk are deemed most appropriate for more focused oversight attention than minimal risk research because they have been identified to entail greater likelihoods of causing harm to human research participants.

2. It identifies those research programs suitable for "expedited review" by an IRB. Research protocols or changes to the protocol requested by the investigator that are reviewed on an expedited basis are not reviewed by the full membership of the IRB, but rather by the IRB director alone. (Research classified as minimal risk need not be reviewed on an expedited basis, however, as an IRB director always has the prerogative to refer a matter to the full board for review.)

3. It identifies those research studies for which the IRB is permitted to allow a modification to the normal requirements for obtaining informed consent. Specifically, minimal risk status is one of the four requirements under Chapter 45, Subpart A, Section 46.116, Paragraph (d) of the *Code of Federal Regulations* that render a research program eligible for an IRB to consider the approval of elimination of one or more of the eight points of informed consent from the consent procedure or even the elimination of the informed consent requirement completely.

Jonathan E. Brill

See also Informed Consent; Institutional Review Board (IRB); Survey Ethics

Further Readings

Department of Health and Human Services. (2005, June 23). *Code of Federal Regulations. Title 45 Public Welfare, Part 46, Protection of Human Subjects*. Retrieved April 14, 2008, from http://www.nihtraining.com/ohsrsite/guidelines/45cfr46.html#46.102

National Bioethics Advisory Commission. (2001, August). *Ethical and policy issues in research involving human subjects, Vol. 1*. Bethesda, MD: National Bioethics Advisory Commission. Retrieved April 14, 2008, from http://www.bioethics.gov/reports/past_commissions/nbac_human_part.pdf

National Institutes of Health, Office of Human Subjects Research. (2007). *IRB protocol review standards*. Retrieved April 14, 2008, from http://ohsr.od.nih.gov/irb/protocol.html

MISREPORTING

Misreporting is the deliberate or nondeliberate reporting of inaccurate or untruthful answers to survey questions. It is often referred to as *response error*. While survey researchers may attempt to gather accurate and truthful responses, respondents are not always willing or able to comply. Misreporting is a major concern for data collected about sensitive topics such as abortion, prejudice, sexual behavior, and income. Misreporting can also occur in the case of nonthreatening questions. Collecting inaccurate and untruthful responses limits the validity of conclusions that can be drawn from survey data.

Respondents may be motivated to deliberately misreport answers when asked sensitive topic questions about behaviors or attitudes for four main reasons: (1) social desirability concerns, (2) protection of the respondent's own self-concept, (3) embarrassment, and (4) fear that unauthorized disclosure may cause harm. When respondents answer questions based on one of these types of motivation, attitudes and behaviors that are socially desirable tend to be overreported (e.g., voting and volunteering), and socially undesirable attitudes and behaviors tend to be underreported (e.g., prejudice and criminal behavior).

Respondents may also misreport answers in a nondeliberate fashion, due to failure in one or all of the following cognitive processes: comprehension of the question, retrieval of relevant memories, judgment about the accuracy and appropriateness of the response, and mapping of relevant information on to a survey response category.

Dennis Dew

See also Cognitive Aspects of Survey Methodology (CASM); Overreporting; Respondent-Related Error; Satisficing; Sensitive Topics; Social Desirability; Underreporting

Further Readings

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

MISSING DATA

An important indicator of data quality is the fraction of missing data. Missing data (also called "item nonresponse") means that for some reason data on particular items or questions are not available for analysis. In practice, many researchers tend to solve this problem by restricting the analysis to complete cases through "listwise" deletion of all cases with missing data on the variables of interest. However, this results in loss of information, and therefore estimates will be less efficient. Furthermore, there is the possibility of systematic differences between units that respond to a particular question and those that do not respond—that is, item nonresponse error. If this is the case, the basic assumptions necessary for analyzing only complete cases are not met, and the analysis results may be severely biased.

Modern strategies to cope with missing data are *imputation* and *direct estimation*. Imputation replaces the missing values with plausible estimates to make the data set complete. Direct estimation means that all available (incomplete) data are analyzed using a maximum likelihood approach. The increasing availability of user-friendly software will undoubtedly stimulate the use of both imputation and direct estimation techniques.

However, a prerequisite for the statistical treatment of missing data is to understand why the data are missing. For instance, a missing value originating from accidentally skipping a question differs from a missing value originating from reluctance of a respondent to reveal sensitive information. Finally, the information that is missing can never be replaced. Thus, the first goal in dealing with missing data is to have none. Prevention is an important step in dealing with missing data. Reduction of item nonresponse will lead to more information in a data set, to more data to investigate patterns of the remaining item nonresponse and select the best corrective treatment, and finally to more data on which to base imputation and a correct analysis.

A Typology of Missing Data

There are several types of missing data patterns, and each pattern can be caused by different factors. The first concern is the randomness or nonrandomness of the missing data.

Missing at Random or Not Missing at Random

A basic distinction is that data are (a) missing completely at random (MCAR), (b) missing at random (MAR), or (c) not missing at random (NMAR). This distinction is important because it refers to quite different processes that require different strategies in data analysis.

Data are MCAR if the missingness of a variable is unrelated to its unknown value and also unrelated to the values of all other variables. An example is inadvertently skipping a question in a questionnaire. When data are missing completely at random, the missing values are a random sample of all values and are not related to any observed or unobserved variable. Thus, results of data analyses will not be biased, because there are no systematic differences between respondents and nonrespondents, and problems that arise are mainly a matter of reduced statistical power. It should be noted that the standard solutions in many statistical packages, those of listwise and pairwise deletion, both assume that the data are MCAR. However, this is a strong and often unrealistic assumption.

When the missingness is related to the observed data but not to the (unknown) value of the missing response itself, it is said that the data are MAR. For example, an elderly respondent may have difficulty recalling an event because of memory problems. The resulting missing datum is related to age but not to the event itself. When the data are missing at random, the missingness is a random process conditional on the observed data. If the data are missing at random *and* if the proper statistical model is used, the missingness is said to be *ignorable* with respect to inference. For example, in the case of the elderly respondent, the variable related to the missingness (age) is measured and available for inclusion in the proper analysis.

Finally, when the missingness is related to the unknown (missing) answer to the question itself, the data are NMAR. For example, a respondent perceives the real answer to a sensitive survey question as socially undesirable (e.g., she or he does have drinking problems) and refuses to respond. If the missing data are the NMAR type, the missingness is said to be *nonignorable*, and no simple solution for treating the missing data exists. A model for NMAR missingness must be postulated and included in the analysis to prevent bias.

Missing Data Patterns

Three main patterns can be discerned in item missing data: (1) the data are missing systematically by design (e.g., contingency questions); (2) all the data are missing after a certain point in the questionnaire (partial completion); and (3) data are missing for some questions for some respondents (item nonresponse).

Missing by Design

Data are missing by design when the researcher has decided that specific questions will not be posed to specific persons. There are two main reasons for items to be missing by design. First, certain questions may not be applicable to all respondents and the questionnaire routing skips these questions for these respondents, that is, these are contingency questions. Since the responses to other questions determine the missingness, the missingness mechanism is accessible to the analyst and can be incorporated in the analyses.

The second reason for items to be missing by design is when a specific design is used to administer different subsets of questions to different persons. In this case, all questions are applicable to all respondents, but for reasons of efficiency not all questions are posed to all respondents. Specific subsets of questions are posed to different groups of respondents, often following a randomized design in an experiment (i.e., random assignment) that makes the missingness mechanism MCAR. Again, since the missingness mechanism is accessible, the incomplete data can be handled statistically and the analyses give unbiased results.

Partial Completion

A partial completion (breakoff) is characterized by time or place dependency. After a certain point in time or place within the questionnaire, *all* data are missing. Partial completions mostly occur in telephone interviews and Web surveys. At a certain time point in the interview, the respondent stops and disconnects. As a result, the remainder of the questionnaire is not answered. When the breakoff occurs early in the questionnaire and only a few questions have been answered, it is usually treated as unit nonresponse. When the breakoff occurs at the end of the questionnaire, the remaining unanswered questions are usually treated as item nonresponse. In that case, information on earlier questions and the interview process is used to investigate the missingness mechanism and adjust for it in the analyses.

Item Nonresponse

Item nonresponse is characterized by blanks in the data for some respondents on some variables. Not every blank in the data matrix originates in the same way. One can distinguish three forms of item nonresponse: (1) the information is not provided by a respondent for a certain question (e.g., a question is overlooked by accident, an answer is not known, a refusal to respond); (2) the information provided by a respondent for a certain question is not usable (e.g., a given answer is not a possible answer, it falls outside the range of permissible responses, multiple responses are given when only one is allowed, it cannot be coded, and/or it is unreadable/illegible); and/or (3) usable information is lost (e.g., error in data entry or data processing). The first two of these mechanisms (information is not provided and information is not usable) originate in the data collection phase. The third is the result of errors in the data processing phase.

The most problematic form of item nonresponse occurs when a respondent does not provide information, because in this case different missing data mechanisms may be at work. When the respondent accidentally overlooks an item, the data are MCAR. The missingness mechanism is ignorable and almost all simple statistical treatments may be used, even listwise deletion. When a respondent is willing but unable to respond—for example, because of memory problems—the missingness depends on an observed variable (age), but not on the answer to the question itself and is thus missing at random. If the data are MAR and if the variable related to the missingness is available, the missingness can be handled adequately with relatively simple solutions. However, when not responding is related to the (unknown) answer to the question itself, the missingness mechanism is NMAR. When a respondent *refuses* to respond, the missingness is probably NMAR and the mechanism is nonignorable. In this case, simple solutions no longer suffice, and an explicit model for the missingness must be included in the analysis.

When item nonresponse is due to unusable responses that are coded as missing, it is generally problematic. The reasons for inadequate responses (e.g., outside the range of possible answers or nonsubstantive responses) are related to the question format and the real value of the answer, pointing to NMAR. If the real answer is partly revealed (e.g., through interviewer notes), the missingness mechanism is at least partly known.

Finally, losing information because of errors in coding, editing, or storing is usually not systematic and therefore normally MCAR. It arises by accident and is not related to questionnaire and respondent characteristics, so the mechanism is ignorable and the solutions are simple.

Analyzing Incomplete Data Sets

Inspecting the Structure and Patterns of Missing Data

For an optimal treatment of item nonresponse, knowledge of the missing data mechanism is valuable. First, one should investigate whether the data are MCAR or not. When incomplete data are MCAR, analyses will not be biased, because there are no systematic differences between respondents who completed the question and respondents who have a missing value for that question.

The first step in the analysis of incomplete data is to inspect the data. This can provide very practical information. For instance, one may find that most of the missing values concern only one specific variable (e.g., household or personal income). But if that variable is not central to the analysis, the researcher may decide to delete it. The same goes for a single respondent with many missing values. In general, however, missing values are scattered throughout the entire data matrix. In that case, a researcher would like to know if the missing data form a pattern and if missingness is related to some of the observed variables. If one discovers a system in the pattern of missingness, one may include that in the statistical analyses or imputation procedures.

The mere inspection of missing data patterns cannot tell the researchers with certainty whether or not the missingness is independent of the (unknown) value of the variable (question). Extra information is needed to test the MAR hypothesis and help to determine the causes of item nonresponse. This information may be available in the data set, but often additional information (information from other sources than the actual sample) is needed, such as theory, logic, or auxiliary data from registers, sampling frames, reinterviews, or other special nonresponse studies.

Effective Methods to Analyze Incomplete Data Sets

The default options of statistical software are usually listwise or pairwise deletion or some simple imputation technique such as *mean substitution*. These solutions are generally inadequate. Listwise deletion removes all units that have at least one missing value and is clearly wasteful because it discards information. Pairwise deletion removes cases only when a variable in a specific calculation is missing. It is less wasteful than listwise deletion, but it can result in inconsistent correlation matrices in multivariate analyses, because different elements in the correlation matrix may be based on different subsamples. Simplistic imputation techniques (e.g., mean substitution) often produce biased point estimates and will always underestimate the true sampling variances. Listwise and pairwise deletion and simple imputation are likely to be biased, because these methods are all based on the strong assumption of MCAR, which seldom is warranted. Therefore, the best policy is to prevent missing data as much as possible, and when they occur to employ an analysis strategy that uses (a) all available information to investigate the missing data patterns and (b) an analysis method that correctly adjusts for missing data.

Only when the data can be considered MCAR do simple solutions like listwise deletion not result in bias. If the fraction of missing data is small, listwise deletion is useful. If the fraction of missing data is large, the MAR-based techniques described following are more efficient.

When the data are assumed MAR, two distinct analysis approaches can be used: direct estimation and imputation.

Direct Estimation

Direct estimation means that the incomplete data are fully analyzed using a maximum likelihood approach. Direct estimation requires specialized software, but this is increasingly becoming available. For instance, several programs for structural equation modeling can include incomplete cases in the analysis. Since analysis of (co)variance, multiple regression analysis, and discriminant analysis can all be formulated as a structural equation model, these analyses can now be done using all available information, under the assumption of MAR. Another example is using multi-level models for incomplete longitudinal data. Such analyses view the repeated measures as hierarchically nested within cases. Since multi-level models do not assume that all measurement occasions are available for analysis, missing data due to panel dropout (attrition) are not a problem.

While direct estimation is powerful, it requires access to and knowledge of specialized software. Imputation fills the gaps in the data set with plausible values, and after the data are made complete, standard software then is used. At this point, the researcher can simply ignore the missingness problem and proceed to analyze the completed data set using any standard method with which she or he is familiar.

Imputation

In imputation, the missing values are replaced by "plausible" values. Many imputation methods exist, which mainly differ in the way they define *plausible*. A problem is that most simple imputation methods, such as replacing missing values with the overall mean or using regression to estimate the missing values, result in biased estimates. However, the popular and reasonably simple *hot-deck* method results in unbiased estimates under the assumption of MAR. In the hot-deck method, the data file is sorted into a number of imputation classes according to a set of auxiliary variables. Missing values are then replaced by observed values taken at random from other respondents in the same imputation class.

There are two fundamental problems associated with imputation. First, using the information in the observed data to predict the missing values emphasizes the structure in the completed data. Second, analyzing the completed data set uses a spuriously high number of cases and thus leads to biased significance tests. Donald Rubin proposes to solve both problems by using *multiple imputation*: Each missing value is replaced by two or more (M) plausible estimates to create M completed data sets. The plausible values must include an error term from an appropriate distribution, which solves the problem of exaggerating the existing structure in the data. Analyzing the M differently completed data sets and combining the estimates into an overall estimate solves the problem of the biased significance test.

In the multiple imputation approach, analyzing M data sets and having to combine the results is cumbersome but not especially complex. What is difficult is generating the M data sets in a proper manner. A nonparametric method is to (a) compute for each respondent the propensity to have missing values on a specific variable, (b) group respondents into imputation classes based on this propensity score, and (c) use hot-deck imputation with these imputation classes. Parametric imputation methods assume a model for the data and use Bayesian methods to generate estimates for the missing values. These methods are described in detail by Joseph L. Schafer. When multiple imputation is used, it is important that the model for the data generation is very general and includes those variables that are important for predicting either missingness or the variables of interest.

Edith D. de Leeuw and Joop Hox

See also Contingency Question; Error of Nonobservation; Hot-Deck Imputation; Ignorable Nonresponse; Imputation; Multiple Imputation; Nonignorable Nonresponse; Nonresponse Error; Panel Attrition; Partial Completion; Random Assignment; Unit Nonresponse

Further Readings

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation*

modeling (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum.
de Leeuw, E. D., Hox, J. J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *15*, 153–176. Retrieved April 14, 2008, from http://www.jos.nu
Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Schafer, J., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavior Research*, *33*, 545–571.

Mitofsky-Waksberg Sampling

Mitofsky-Waksberg sampling is a two-stage, clustered approach for selecting a random sample of telephone numbers. Developed by Warren Mitofsky and Joseph Waksberg in the 1970s, this was an innovative approach designed to improve the operational efficiency of telephone samples through reductions in the proportion of unproductive numbers dialed. Prior to the development of Mitofsky-Waksberg sampling, unrestricted random-digit dial (RDD) was used, but this method was operationally inefficient as it led interviewers to call far too many nonworking numbers. Mitofsky-Waksberg sampling (including modified versions of the basic approach) was the predominant approach used for selecting samples for RDD telephone surveys throughout the 1970s and 1980s, but it was largely supplanted by list-assisted RDD by the early 1990s.

An understanding of the various approaches used for RDD sampling requires some knowledge of the structure of a telephone number. In the United States, telephone numbers are 10-digit strings. The first three digits are the *area code*, and the first six digits are the *telephone exchange*. A *100-bank* is a set of telephone numbers having the same first eight digits. Historically, telephone numbers were geographically clustered. However, under provisions of the Telecommunications Act of 1996, customers are able to retain their telephone numbers when switching from one telephone service provider to another, even when that switch involves a geographic move or a switch between landline service and cellular service. This is called "porting" a telephone number.

Telephone exchanges are designated for particular uses (e.g., cellular only, plain old telephone service [POTS] only, cellular and paging). For RDD surveys, the sampling frame of telephone exchanges has typically been based on those exchanges that are designated for POTS (i.e., landline) use. However, within that subset of exchanges, not every number is assigned, and not every assigned number is residential. Thus, efficiency gains may be achieved by reducing the number of unproductive (nonworking or nonresidential) telephone numbers that are dialed.

Implementation

In the first stage of selection in the Mitofsky-Waksberg approach, the set of telephone exchanges is limited to those exchanges designated for residential use, and a sample of 100-banks is selected for the sampling frame. A random two-digit suffix is appended to each sampled 100-bank to obtain the *prime number*. Each prime number in the sample is dialed to determine whether it is a residential number. If the prime number is a residential number, the 100-bank is retained in the sample, and in the second stage of selection, additional telephone numbers (*secondary numbers*) are selected in that 100-bank. If the prime number is not a residential number, then the 100-bank is excluded from the second stage of sampling. Following the second stage of selection, attempts are made to complete interviews until a predetermined fixed number (k) of residential numbers is identified among the secondary numbers in the 100-bank. The total number of residential numbers in the sample is $m(k+1)$.

A disadvantage of the Mitofsky-Waksberg method is that the selection is sequential; all primary numbers must be resolved before the second stage of sampling can occur, and each secondary unit must be resolved before additional units can be selected. Noncontact cases (ring–no answer and answering machine results) are problematic in that regard. Richard Potthoff, J. Michael Brick, and Joseph Waksberg each developed modified Mitofsky-Waskberg methods to address the sequential nature of the sample.

Efficiency

To evaluate the efficiency of the Mitofsky-Waksberg approach, the precision of survey estimates and the cost of the approach relative to unrestricted RDD are discussed following.

Effect on Precision

Let m denote the number of 100-banks in the sample, σ^2 the unit variance of a characteristic y, and ρ (rho) the intraclass correlation of the characteristic y (i.e., the correlation in y among units in the same 100-bank). Mitofsky-Waksberg sampling results in an equal probability sample of residential telephone numbers. Therefore, the effect of using this approach on the variances of survey estimates is due to clustering of the sample of telephone numbers within exchanges. The variance of the sample mean $\bar{y}$ is approximately $V_1 = \frac{\sigma^2}{m(k+1)}[1+k\rho]$, and, therefore, the effective sample size is $\frac{m(k+1)}{1+k\rho}$.

Effect on Cost

Let π denote the proportion of eligible residential numbers in the population, and let t denote the proportion of 100-banks with no eligible residential numbers. Further, let C_p/C_u denote the ratio of the cost of a productive call to the cost of an unproductive call. Then the cost of Mitofsky-Waksberg sampling, relative to unrestricted RDD, is given by

$$\frac{C_{M-W}}{C_{unrest}} = \frac{\frac{C_p}{C_u} + \frac{1}{\pi}\left[1 - \pi - t\left(\frac{k}{k+1}\right)\right]}{\frac{C_p}{C_u} + \frac{1}{\pi}[1-\pi]}(1+k\rho).$$

Advantages and Disadvantages

Historically, the benefit of Mitofsky-Waksberg sampling was that, relative to unrestricted RDD, it greatly reduced the number of unproductive numbers dialed. A disadvantage of Mitofsky-Waksberg sampling is that it results in a clustered sample, and thus it results in some loss of precision due to intraclass correlation. A second disadvantage is the sequential nature of the selection. It was an innovative approach that served as the predominant sampling method for telephone surveys for more than a decade. However, due to its shortcomings, Mitofsky-Waksberg sampling has been replaced by list-assisted RDD.

Jill M. Montaquila

See also Cluster Sample; Effective Sample Size; EPSEM Sample; List-Assisted Sampling; Number Portability; Random-Digit Dialing (RDD); ρ (Rho); Sequential Sampling; Telephone Surveys

Further Readings

Brick, J. M., & Waksberg, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, *17*, 27–41.

Potthoff, R. F. (1987). Some generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, *82*, 409–418.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, *73*, 40–46.

Mixed-Mode

Mixed-mode surveys (sometimes referred to as *multi-mode surveys*) combine different ways (modes) of collecting data for a single project. Different methodologies may be used during distinct phases of a survey, such as recruitment, screening, and questionnaire administration, or they may make use of different survey modes during a single phase, like data collection. Mixed-mode surveys may involve combinations of more traditional survey modes such as face to face, telephone, and mail, or may include some of the newer modes like Internet, cell phone, diaries, or interactive voice response (IVR).

Reasons for Use

Researchers often employ mixed-mode survey designs to address problems associated with undercoverage of key groups of interest or to improve participation rates. Some mixed-mode designs can also be conducted at lower cost than single mode surveys, particularly when survey modes costing less (e.g., mail) are employed first to obtain a large percentage of the completed interviews and more costly methods (e.g., in person) are used later to attempt to interview initial nonrespondents.

The decision to use a mixed-mode approach and the particular design to employ is based on several considerations. First is the need to identify the survey design that best meets the study's objectives, in terms of the research question(s) posed, population of interest, and amount of data to be collected. Next is the desire to reduce the total survey error in a project, which is the error from all potential sources, including coverage, sampling, nonresponse, and measurement error. The decision is also affected by the time frame available for data collection. Some modes (such as mail surveys) require considerably longer field periods than other modes (such as telephone surveys). Finally, cost is an important consideration, given that researchers typically need to operate within a fixed budget.

Bias and Cost Differences

From a research perspective, an optimal mixed-mode design is one that for a fixed budget will reduce the total amount of error in the survey to the greatest extent possible, thereby reducing the potential for bias in the survey estimates. This typically involves consideration of survey coverage, sampling error, nonresponse, and measurement error. The frame used for sampling potential respondents must include all units in the population of interest. If particular types of sample units (persons, households, businesses, etc.) are not included in the sampling frame, then coverage bias may result. Modes differ in their degree of coverage, with face-to-face surveys (based on residential addresses) often having the highest levels of coverage and Internet the lowest.

Coverage rates by mode can and do change over time. For instance, in the late 1990s landline telephone surveys had household coverage rates of 95%–98%. Since the advent and rapid adoption of cellular telephones, however, landline telephone frame coverage of households has dropped significantly (e.g., to less than 80% by 2008). Coverage of the general population for mail survey has improved steadily over time with the development of computer technology and large address databases. Internet surveys of the general population based on sampling email addresses tend to have very low coverage because there is no centrally maintained or comprehensive listing of email addresses. However, Internet surveying via email addresses of known lists of individuals (e.g., members of a professional organization or students at a university) can have extremely high coverage.

In terms of sampling in a mixed-mode survey, each sampled unit should have a known and nonzero (but not necessarily equal) probability of selection. Face-to-face and telephone surveys have a variety of well-established methodologies for sampling units from frames. Within-household selection techniques are less well developed for mail surveys of the general population. With some exceptions, Internet-based surveys of the general public often tend to be nonprobability surveys by which respondents decide to opt in to the survey rather than being chosen to participate through a random selection process.

Nonresponse results when a researcher is unable to obtain data from a sampled respondent. Unit nonresponse, where the selected household or person does not participate in the survey at all, tends to be lowest in face-to-face surveys and highest in Internet surveys. Item nonresponse, where a selected respondent chooses not to answer a particular survey question, tends to vary by survey mode and question topic.

Measurement error occurs when a respondent's answer to a question is inaccurate, that is, it departs from the "true" value. Measurement error can vary across survey modes depending on whether the survey is conducted by an interviewer or is self-administered. When an interviewer administers a survey, he or she can help to motivate the respondent, guide the respondent through complex questionnaires, clarify questions or instructions, and probe for more detailed responses when necessary. Interviewer-administered surveys can, however, offer respondents less privacy and anonymity, lead to socially desirable and more positive responses, and result in acquiescence (when a respondent goes along with what the respondent thinks the interviewer wants to hear) or social desirability (when the respondent answers questions in a way that puts herself or himself in a good light in the eyes of the interviewer).

In contrast, self-administered modes can often ensure greater privacy, let respondents proceed through the questionnaire at their own pace, and allow respondents to complete the survey at a time convenient to them. The drawbacks of self-administered modes often include a loss of control by the researcher in ensuring the correct respondent completes the entire questionnaire (as is the case with a mail survey), an increase in stray or out-of-range responses (when using noncomputerized questionnaire applications), and no means of assessing the level of cognitive engagement of the respondent in the survey (that is, the degree to which a respondent is answering the questionnaire in a serious and thoughtful manner).

Measurement error can also result from the type of stimuli or manner in which the survey question is conveyed to the respondent. A survey question may be presented visually (such as on a paper questionnaire or a computer screen) or aurally (such as when an interviewer reads a question to a respondent). Respondents have been shown to answer the same question differently when a question is read versus when it is verbally asked. Likewise, the manner in which a respondent provides the answer—spoken, written, or typed—can affect the response, including the amount of time given to the response, the level of thought involved, and the amount of detail provided.

Modes also vary in terms of their cost per completed interview. While there is wide variation by design, typically face-to-face surveys are the most expensive to conduct, while Internet surveys are the least costly. Telephone and mail surveys tend to fall between these two extremes. It is important to remember, however, that the per-unit costs may be high even for a traditionally low-cost mode if the mode is seldom used by respondents. Generally, the more interviews completed using a particular mode, the lower the per-unit cost will be for that mode.

Considerations for Combining Modes

Operationally there are a number of issues a researcher needs to consider when combining survey modes in a mixed-mode survey. First is reaching the population of interest. The population of interest needs to be reached by the combination of modes being employed. To this end, researchers need to understand key elements of the population they are trying to reach, including their physical accessibility, telephone access (landline or cell phone), literacy level, and access to the Internet. Additionally, researchers need to consider how particular subgroups of the population might want to be contacted and respond to a survey. For instance, the Internet or cell phones may not be a good way of reaching a population of individuals ages 65 years or older, given that Internet and cell phone usage among this group is relatively low compared to other age groups. Likewise, use of an English-language-only questionnaire might not be the best match for a population of recent immigrants.

Second, a determination needs to be made as to whether the modes will be used sequentially or concurrently. For sequential assignment, different modes can be used for successive phases of a survey (contacting, screening, and data collection) or used sequentially during the data collection phase. For cost efficiency, it is typically better to use the least expensive mode(s) (e.g., mail and/or Internet) first, followed by the more expensive mode(s) (telephone and/or face to face). The sequential approach also provides the researcher with greater control over the survey process. In contrast, concurrent assignment provides multiple channels of contact to be used simultaneously. Typically, the respondent is allowed to choose the mode that best suits her or his needs.

Third, researchers need to be cognizant of the potential limits on comparability across modes. For instance, changing modes at different points in time of a longitudinal survey or panel survey may lead to a confounding of time effects (differences in responses due to changes in responses over time) and mode effects (differences in responses resulting from a difference of survey mode). Similarly, if different modes are used across different subgroups of the population, it may become difficult to distinguish between real differences in survey responses among these subgroups and differences due to mode effects.

Fourth, in using mixed-mode approaches, researchers should strive during the design phase of the project to reduce the potential for measurement error wherever possible. Modes can differ in terms of the format in which a question is presented (for instance, interview-administered surveys tend to present one question at a time, whereas mail and Internet surveys will often allow respondents to see blocks of questions before providing a response).

Don Dillman, one of the foremost experts in the topic of mixed-mode surveys, suggests the use of a *unimode design*, which consists of writing and presenting questions in a way that assures receipt by respondents of a common mental stimulus regardless of survey mode. Such a design would involve (a) making all response options the same across modes and incorporating them into the stem of the survey question; (b) avoiding changes in the basic question structure across modes that could change the stimulus; (c) reducing the number of response categories to achieve mode similarity; (d) using the same descriptive labels for response categories instead of depending on respondents' vision to convey the nature of the scale concept; (e) developing equivalent instructions for skip patterns that are determined by the answers to several widely separated items; (f) avoiding question structures that unfold; (g) reversing the order in which categories are listed in half of the questionnaires; and (h) evaluating interviewer instructions carefully for unintended response effects.

Finally, researchers need to consider several other factors. Mixed-mode surveys require a greater level of in-house expertise; that is, researchers need to fully understand the strengths and limitations of the modes they propose to use and combine. Implementation and timing for a mixed-mode survey is often more complex and takes longer than the average single-mode survey (depending on which modes are involved). There are data processing considerations, in terms of combining and weighting or adjusting the data from a mixed-mode survey. The quality control requirements can often vary by mode. Finally, researchers need to develop a way of collecting, combining, and assessing paradata (i.e., operational data such as call counts, case dispositions, the days and times a household was visited, etc.) about the mixed-mode survey to ensure integration across the modes used.

Michael W. Link

See also Address-Based Sampling; Aural Communication; Coverage; Coverage Error; Face-to-Face Interviewing; Field Period; Interactive Voice Response (IVR); Internet Survey; Mail Survey; Measurement Error; Mode; Mode Effects; Mode of Data Collection; Mode-Related Error; Multi-Mode Surveys; Nonresponse; Nonresponse Error; Paradata; Probability of Selection; Sampling Error; Survey Costs; Telephone Survey; Total Survey Error (TSE); Undercoverage; Visual Communication; Web Survey

Further Readings

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

de Leeuw, E. (2005). To mix or not to mix data collection methods in surveys. *Journal of Official Statistics*, *21*, 233–255.

Dillman, D. A. (2006). *Mail and Internet surveys: The tailored design method*. New York: Wiley.

Link, M. W., & Mokdad, A. H. (2006). Can Web and mail survey modes improve participation in an RDD-based national health surveillance? *Journal of Official Statistics*, *22*, 293–312.

Mode

The mode is a type of descriptive statistic that researchers commonly use to characterize the data from their studies. Along with the *mean* (average) and *median*, the *mode* constitutes one of the measures of central tendency—a general term for a set of values or measurements located at or near the middle of the data set. The mode is calculated as the most frequently occurring value within a set of observations.

For example, in a data set containing the values 1, 1, 2, 3, 4, 5, 6, 7, 8, and 9, the mode would be the value 1, as it is the value within the data set that appears most often. However, a data set can have more than one mode, in which case, it is bimodal or even multi-modal. For instance, in a data set containing the values 1, 1, 2, 2, 3, 4, 5, 6, 7, 8, and 9, the modes would be 1 and 2, as they appear the most often.

The mode is commonly used to measure the most popular value among a set of categorical values. For instance, in a response to a survey question that has four choices : A (selected 15% of the time), B (50%), C (15%), or D (20%), the mode would represent the most popular choice among the four choices A through D. In this example, the mode would be B, with 50% of the selected values. The mode can also be used with other data scales (ordinal, interval, ratio), but researchers should be careful to select the appropriate metric to best represent the data available. Depending on whether the data are distributed uniformly in a normal distribution (bell-shaped curve) or skewed in one direction or another, the mode may or may not be equivalent (or even close in value) to the mean or median.

Richard Kwok

See also Mean; Median

Further Readings

Kornegay, C. (1999). *Math dictionary with solutions: A math review*. Thousand Oaks, CA: Sage.

Mode Effects

Survey researchers use the term *mode* to refer to the way in which data are collected in the survey. Often, mode will be used to refer specifically to the way the questionnaire is administered (e.g., as a self-administered paper-and-pencil questionnaire, on the Internet, or as a face-to-face interview). However, mode can be discussed as a facet of various phases of a survey project, including sampling, contact, and recruitment, as well as the format of the questionnaire itself. One area of survey research and methodology looks at the impact of mode on data obtained from surveys. This area of research is generally called "the study of mode effects." At its most general, the term, *mode effects*, refers to any influence on survey responses that is due to the mode of data collection. It quickly becomes clear that this definition encompasses a large body of phenomena.

Components of Mode and Causes of Mode Effects

When considering mode effects, it can be helpful to think about the social, psychological, physiological, and technical facets that comprise a given mode. For example, a face-to-face interview mode usually involves a one-on-one social interaction between a respondent and an interviewer, which generally carries with it certain social norms. Physiologically, the respondents must have the ability to hear survey questions if they are to be presented verbally by an interviewer. The interviewer and respondent must also be able to converse in a common language. Finally, there are a number of logistical issues surrounding traveling to and from sampled persons and finding a place to conduct the interview.

Interviewer Presence

Mode differences can be thought of in terms of dimensions on which modes differ. One of these dimensions is the degree of interviewer involvement. A purely self-administered mode (e.g., a paper-and-pencil survey that is mailed to respondents) removes this component completely.

The characteristics of the interviewer and respondent can impact responses to some kinds of questions in studies in which an interviewer is involved in the data collection. For example, it has been found that the match between the race of the interviewer and the race of the respondent can influence responses to racial attitude questions. More moderate answers have been found to be reported by African American respondents to white interviewers than to African American interviewers. Similar results have been found for the gender of the interviewer. The match of the gender of the interviewer and respondent may be important, as more "feminist" answers are reported to female interviewers.

Interviewer variance is another type of interviewer-related mode effect that arises when responses of respondents interviewed by the same interviewer tend to be correlated with each other. This phenomenon is found in both forms of interviewer-administered surveys (face-to-face surveys and telephone surveys), and it contributes to the variance component of statistical error. Interviewer variance, noted by ρ_{int}, is a specific application of the intraclass correlation. If $\rho_{int} = 1.0$, then responses within an interviewer's set of respondents are completely correlated. If it is zero, they are completely uncorrelated. The study of interviewer variance requires an interpenetrated design in which interviewers are randomly assigned to respondents so that natural intraclass correlation (e.g., those due to neighborhood or region) can be separated from intraclass correlation caused by the interviewer. For this reason, few valid studies have been done. Nonetheless, ρ_{int} values tend to be higher for attitude questions than for factual questions. They are also found in open-ended questions, when interviewers need to probe in order to get further responses. The absolute values of ρ_{int} tend to be fairly small, though they have been found to be larger in face-to-face interviews than in phone interviews. However, the absolute value is of less concern than the impact on the error of a statistic. The impact of ρ_{int} on a given statistic is determined by the size of the interviewer's workload. The impact of interviewer variance on a statistic is $[1 + (\bar{m} - 1) \times \rho_{int}]$, where $\bar{m}$ is the average interviewer workload. This form of mode-related variance is not found in self-administered data collection modes.

Contact With Respondent

A second dimension in which modes differ is the degree of contact with respondent. Even though interviewers are involved in both face-to-face and telephone interviews, they have very different levels of contact with the respondents, specifically in the "distance" created by the telephone in that mode. A respondent and interviewer sitting across from each other in the same physical space may create a different social psychological context than that created by the interviewer–respondent interaction over the telephone.

Privacy

A third aspect of survey mode that is directly related to interviewer involvement is privacy. Particularly when topics are sensitive or personally revealing (such as disease status, sexual orientation, or income), the respondent's perception of privacy may be an important factor in her or his decision to report information accurately or to participate in the interview at all. With respect to mode, some of the most robust findings about privacy indicate that self-administered questionnaires produce fewer socially desirable responses than questionnaires involving an interviewer. This is particularly the case for sensitive behaviors, such as sexual practices or the use of illegal substances. In many cases, higher reports of sensitive behaviors are taken to be more accurate (when a comparison with another data source is not available). However, with sexual activity, socially desirability effects seem to work in opposite directions for men and women, with women reporting fewer sex partners to an interviewer and men reporting more. It is clear that the presence of an interviewer can produce mode-related measurement error, but the direction of that error is not always clear.

Channels of Communication

Fourth, channels of communication will differ by mode of administration. A telephone survey requires aural and verbal channels. A self-administered questionnaire generally requires only visual channels, but some also include aural channels. Not all communication channels may be necessary for the survey task. For example, turning a telephone survey into a face-to-face survey significantly increases the number of communication channels available (from sound only to visual). This can be beneficial to the researcher who decides to use these channels in the survey protocol, by using show cards with response choices in a face-to-face interview, for example. But they may also backfire in the sense that the additional channels of communication may provide information that is not relevant to the survey task but still influences answers, as in the case of a social desirability-prone respondent in the presence of an interviewer. The face-to-face dynamic brings along with it nonverbal communication, visual appearance, and other facets that may have an impact on nonresponse and measurement error.

Technology Use and Ability

Finally, technology usage is a dimension on which survey modes can vary. How much technology does the researcher use in the data collection protocol? How comfortable are the respondents with the particular technology? There are currently a number of different applications and variations of computer-assisted interviewing (CAI), some of which retain interviewers as part of the data collection, as in computer-assisted personal interviews (CAPI) or computer-assisted telephone interviews (CATI), and others that allow respondents to administer the questionnaire themselves, such as audio computer-assisted self interviewing (ACASI), interactive voice response (IVR), and Web surveys. Web surveys are a type of computerized data collection that has grown in popularity over the past decade, primarily due to their low cost and relative ease of implementation. In this type of computerized data collection, the respondent interacts directly with the technology, and so the ability to use the technology is not only an issue of design and management from the researcher's point of view but is also an issue of respondent acceptance of and ability to employ the technology.

Understanding Mode Effects Research

Due to logical constraints on survey research, modes are often tied to packages of survey protocols (e.g., sample selection, recruitment, measurement) in ways that make the most sense for operations and cost. At the same time, the nonresponse and measurement aspects of mode can be dissected into component processes that may lead to effects, for example, the social dynamic involved in socially desirable responding. In reading research on mode effects, one needs to be aware of whether the comparison being done involves a "mode package" that would be found in typical survey practice or a manipulation of the specific mode components, which might be found in more theoretically oriented survey research.

Sampling and Mode

Several phases of the survey process can be impacted by mode. Sampling is directly affected by mode as the selection of an appropriate sampling frame is related to the mode. Does the researcher purchase a sample of telephone numbers for a phone survey or design an area probability sample from census data and maps for face-to-face interviews? In reality, it may be more likely that one's measurement mode is influenced by the sample available or the sample required for the inference needed (e.g., for national inference to the general population, telephone or area probability samples are generally needed). Sampling error deals specifically with the sampling and estimation decisions made by analysts and is primarily a statistical problem, which is why mode effects are not clearly and directly related to sampling per se. However, there are mode effects in coverage, which is a part of the sampling and inference process.

Coverage and Mode

Coverage error in surveys results when there are differences between respondents who are included in the sampling frame (i.e., covered) and those who are in the target population but are not included in the frame. An example of coverage problem related to mode is the use of Web surveys when one wants to produce findings representative of the general public. According to a 2005 study by researchers at the Bureau of Labor Statistics and Westat, 89% of the U.S. households surveyed had a landline telephone (and would thus be covered by an RDD frame). Another 6% only had cell phones, and 5% had no phone. In comparison, in 2003, according to the U.S. Census Bureau, 62% of households had a computer, and 55% of households had access to the Internet. If having access to the Internet at home is the requirement for responding to a Web survey, which is most likely the case for individuals who do not have Internet access at work or school, there would be an noncoverage rate of 45% for Internet surveys in the United States as of 2003. This would be a minimum undercoverage that would be accentuated by any additional sampling based on specific information about Internet uses, that is, America OnLine subscribers, emails with certain domain names, and the like. Furthermore, only about 20% have high-speed access, which may be needed for certain Web survey applications.

If these coverage rates were randomly distributed across the population, one would not need to worry so much about the potential for coverage error. However, there are differences by age and education in the coverage of a Web survey frame, with older individuals and less-educated individuals being far less likely to be covered. In addition, the 95% coverage rate of households with telephones does not necessarily indicate the absence of mode-related coverage error. In this group, the most likely not to be covered are the extremely poor. So, a survey of poverty that used a telephone sample and interview mode may not be a prudent decision. Similarly, other variables related to extreme poverty (e.g., illiteracy, homelessness, and substance abuse) are not best measured through this mode due to coverage problems.

There are some data that show differences in Internet access and use by occupation, education, and gender. In an analysis of computer and Internet access patterns over time (1995 to 2002), it was found that gaps in access by occupation, education, and gender diminished, but highly educated males still were most likely to have computer access and Internet use both at home and work. An interesting pattern is that while gaps in access to computers and the Internet have decreased, the gaps in the amount of time spent using them have increased. In terms of mode differences, this has implications for coverage. If coverage is defined as any access to the Internet, the occupation, education, and gender differences may be acceptable. However, if coverage is defined as "all individuals who have access to the Internet at a single point in time" (e.g., 9:00 p.m. on January 12, 2008), then there is a problem with undercoverage of older individuals and those with less education, since those people tend to be online less often. In terms of mode-related nonresponse due to noncontact, less time on the Internet may mean less chance of reading and responding to an email request for survey participation.

There is also potential for mode effects in coverage in cases where an interviewer is involved in the sampling process in some way. For example, many survey samples involve selection of households, followed by the listing of household members, and then selection of an individual member within the household to answer the survey questions. If the interviewer errs in taking the household listing and in selecting a respondent, or if something about the process leads respondents to answer the listing questions incorrectly, then mode-related within-unit coverage error can result. In a recent U.S. Census, an undercount of young African American males was noticed, and further research has suggested that it was due to the household listing process conducted by interviewers.

Unit Nonresponse and Mode

Assuming that individuals are in the sampling frame and have a nonzero probability of selection (i.e., they are covered), modes can affect nonresponse by influencing contact likelihood and other aspects of unit-level nonresponse. Contact involves the ability to present the survey request to the selected individual or household, while other types of unit nonresponse (given contact) can include outright refusals to the survey, refusals for logistical or illness reasons (too busy or too ill during the field period), or default refusal by continually ignoring approaches by the researcher (i.e., never returning calls or answering the door). Contact may seem like a nominal issue, but for certain segments of the population, in-person contact, which is needed for a face-to-face survey, can be extremely difficult (e.g., because of gated communities in wealthier neighborhoods or high crime rates in poor neighborhoods).

Similarly, clear differences in response rates can be seen by mode. Response rates are generally higher in mail surveys than in Web surveys, with a few exceptions. Exceptions are likely due to differences in the target population, the recruitment procedures used, or the quality of the list or procedures used to contact respondents. Further, face-to-face surveys obtain higher response rates on average than telephone, mail, or Web surveys. This is thought to be due in part to the in-person presence of an interviewer who can tailor the approach to the potential respondent based on verbally expressed concerns, nonverbal behavior, and other contextual factors such as the presence of children, the smell of cooking, or the like.

Item Nonresponse, Measurement Error, and Mode

The mode of survey administration can also have an impact on survey results at the level of the survey question (item). This includes both item nonresponse (that is, missing data) and measurement error on items that have been reported.

Item Nonresponse

Just as potential respondents can refuse to participate in a survey, those who agree to participate can choose not to answer individual questions. Such phenomena are relatively rare in "average" survey questions (below 5%), but can be fairly high for certain kinds of questions like income (upward of 30% to 40% in general population surveys). For most types of survey questions, face-to-face and telephone surveys produce much lower rates of missing data. The cause is thought to be related to the presence of the interviewer, specifically the task of making sure that every question is asked, thereby eliminating or at least vastly reducing item nonresponse due to respondents inadvertently skipping or refusing a question. That is, it is more difficult for most respondents to refuse to answer a question asked by a live person (the interviewer) than a question asked through the medium of paper or a computer screen.

Measurement Error

Mode-related measurement error can be linked to the respondent, the interviewer (if relevant), or the instrument or questionnaire. Each of these facets varies in presence, level, or quality by the mode of data collection. A measurement error occurs when a respondent answers a question but does so with an answer that is not their "true score" for that question. This could be an endorsement or denial of a specific behavior or opinion that is different from their actual behavior or attitude. It can also occur on continuous or pseudo-continuous variables and result in an overreport or underreport of a behavior or strength of an attitude.

Sometimes a nontypical measurement mode, like a diary, is the best way to get at a certain behavior or characteristic. For example, when respondents are asked to keep a diary about drinking behavior and then are also asked to recall their drinking behavior over a specified time frame, diaries are found to be more accurate. In this example, the mechanism explaining the mode difference is likely to be the fallibility of human memory for past behaviors and events.

Primacy and recency effects involve the likelihood that respondents will select items that fall earlier (primacy) or later (recency) in a list of response options. The effects are thought to be due to the cognitive and perceptual processes inherent in different modes. In a self-administered mode, respondents can, in theory, consider all the response options at a glance and pick the one most appropriate for them. However, there is a well-documented tendency for a respondent to pick the best *first* response that they encounter, and thus a primacy effect results. In interviewer-administered modes where respondents hear only the response choices, there is a tendency to observe recency effects. After the interviewer has read all the response options, respondents are more likely to answer the question by choosing options toward the end of the list that they heard most recently with a higher probability than those at the beginning of the list. This is thought to be due to the role of short-term or working memory in retaining all the potential options. Respondents may be more likely to pick more recently heard options because they are the only ones remembered. Sometimes primacy and recency effects are lumped into a general family of nonoptimal responding called *satisficing*, but it seems reasonable to think that they may also be due to cognitive and perceptual limitations.

Cost and Multi-Mode Options

A final mode effect to consider is cost. Face-to-face interviews are clearly the most expensive form of data collection, due heavily to costs of travel and other tasks related directly to the mode. Telephone surveys are less expensive than in-person surveys due to factors such as the lack of travel costs, the centralization of staff, and quicker and less costly transition from one interview to another. Mail surveys can be fairly inexpensive, requiring only the costs of printing, mailing, and data processing; but costs increase depending on follow-up mailing to nonresponders. Web surveys may be the cheapest form of data collection, since they remove data entry (but not the need to edit or code data), printing, and mailing costs associated with mail surveys, as well as interviewing staff. Follow-ups in Web surveys do not affect costs as they do in mail surveys.

Mixed-Mode Surveys

Considering all the variations of survey modes, the impacts of effects and costs, there seem to be benefits and drawbacks for different modes given the specific statistics or measures needed. Survey researchers have begun to use multiple modes in single surveys as ways to counterbalance mode effects as sources of error by building on the strengths of one mode to offset the limitations of another. For example, an in-person nonresponse follow-up with nonresponders to a mail survey is one example. Using an interviewer-administered mode for most of the survey questions but a self-administered mode for sensitive questions is another.

Matthew Jans

See also Audio Computer-Assisted Self-Interviewing (ACASI); Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Diary; Interactive Voice Response (IVR); Interviewer Effects; Interviewer-Related Error; Interviewer Variance; Mixed-Mode; Mode of Data Collection; Mode-Related Error; Multi-Mode Surveys; Primacy Effects; Recency Effects; Satisficing; Self-Administered Questionnaire; True Value; Web Survey

Further Readings

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

de Leeuw, E. D., & van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: A comparative meta-analysis. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283–300). Hoboken, NJ: Wiley.

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Groves, R. M, Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews*. New York: Academic Press.

Losch, S. C. (2004). Gender, education and occupational digital gaps 1983–2002. *Social Science Computer Review*, *22*, 152.

Schober, M. F., & Conrad, F. G. (Eds.). (2008). *Envisioning the future of the survey interview*. Hoboken, NJ: Wiley.

Tourangeau, R., Rips, L. J., & Rasinksi, K. (2000). *The psychology of survey response* (Chapter 10). Cambridge, UK: Cambridge University Press.

Tucker, C., Brick, J. M., & Meekins, B. (2007). Household telephone service and usage patterns in the United States in 2004: Implications for telephone samples. *Public Opinion Quarterly*, *71*(1), 3–22.

U.S. Census Bureau. (2003, October). *Computer and Internet use in the United States: 2003*. Retrieved June 18, 2007, from http://www.census.gov/population/www/socdemo/computer.html

Model-Based Estimation

The primary goal of survey sampling is the accurate estimation of totals, means, and ratios for characteristics of interest within a finite population. Rather than assuming that sample observations are realizations of random variables satisfying some model, it is standard to treat only the sample selection process itself as random. This is called *randomization* or *design-based inference*. Because they rely on averages taken across all possible samples and not on the sample actually drawn, design-based methods can sometimes produce misleading results. *Model-based estimation*, by contrast, is conditioned on the realized sample but requires more assumptions about the behavior of the characteristics of interest. Model-based methods can be used along with or as a substitute for design-based inference.

Let U denote the population of N elements. Suppose the goal is to estimate the total y value for U, that is, $T = \sum_{k \in U} y_k = \sum_{k=1}^{N} y_k$, from a sample S of $n < n$ elements (observations). Under a simple model in which the y_k are uncorrelated random variables with a common mean, say μ, the estimator $t = (N/n) \sum_{k \in S} y_k$ is an unbiased estimator for T in the sense that $E_M(t - T) = E_M(t) - E_M(T) = N\mu - N\mu = 0$, where the subscript M indicates that the expectation is with respect to the model. One needs to add the assumption that the sampling mechanism is such that the expected value of y_k for elements within the sample is the same as that for those outside of it.

The estimator t is identical to a standard estimator for T under design-based theory *when the sample is selected using simple random sampling without replacement*. Thus, a mild restriction on how the sample can be chosen allows one to make a valid inference in some sense without resorting to any model assumptions at all. Why then would anyone use model-based estimation?

Consider this common situation. Suppose for every element $k \in U$, one knows the value of an auxiliary variable, x_k, believed to be closely related to y_k. For example, k can be a high school in state U, x_k an administrative record of the number of 12th graders in high school k, and y_k is the number of 12th graders in high school k applying for college as measured by a complete enumeration of the 12th graders in the school.

It is often not unreasonable to assume that the y values in U obey the ratio model:

$$y_k = \beta x_k + \varepsilon_k,$$

where the $\varepsilon_k | x_k$ (i.e., ε_k given x_k) are uncorrelated random variables with mean zero. Given a simple random sample of size n, t will be model unbiased for T only when $E_M(t) = {}^{N}/_{n} \sum_{k \in S} \beta x_k$ equals $E_M(T) = \sum_{k \in U} \beta x_k$ or, equivalently, when the sample mean of the $x_k, \bar{x}_S = \frac{1}{n} \sum_{k \in S} x_k$, equals the population mean, $\bar{x}_U = \frac{1}{N} \sum_{k \in U} x_k$. This happens *on average* across all possible simple random samples but will not usually happen for a *particular* selected sample.

When $\varepsilon_k | x_k$ has the same distribution whether or not k is in the sample, the sampling mechanism is said to be *ignorable*. Given a sample selected with an ignorable mechanism, a model unbiased estimator for T is

$$t_{ratio} = \left(\sum_{k \in U} x_k \right) \left(\sum_{k \in S} y_k \Big/ \sum_{k \in S} x_k \right).$$

If $E\left(\varepsilon_k^2 | x_k\right) \propto x_k$ (i.e., the variance of ε_k given x_k is proportional to x_k) for all $k \in U$, then one can show that the sample minimizing the model variance of t_{ratio} as an estimator for T is the *cutoff sample* containing the n elements in U with the largest x values. One does not even have to add the assumption that $\varepsilon_k | x_k$ has the same distribution within and outside the sample since the random variable is defined conditioned on the size of x_k, which is the only criterion used in cutoff sampling.

Many surveys designed to measure change are based on either cutoff samples or samples selected for convenience. In this context, x_k is a previous value known for all elements in U, and y_k a current value known only for elements in S. The ratio model and the ignorability of the sampling mechanism is assumed (perhaps only implicitly), and t_{ratio} is computed.

When the sampling mechanism is ignorable, there are many unbiased estimators for T under the ratio model. Some are more model efficient (have less model variance) than t_{ratio} when $\sigma_k^2 = E\left(\varepsilon_k^2 | x_k\right)$ is not proportional to x_k. Usually, however, assumptions about the σ_k^2 are on less firm ground than the ratio model to which it is attached. Moreover, the model itself, although apparently reasonable in many situations, may fail because the expectation of the $\varepsilon_k | x_k$ subtly increases or decreases with the size of x_k.

Design-based methods offer protection against possible model failure and the nonignorability of the sampling mechanism. These methods, however, often depend on a different kind of assumption—that the realized sample is sufficiently large for estimators to be approximately normal. Combining design and model-based methods is often a prudent policy, especially when samples are not very large. A working model that is little more than a rough approximation of the stochastic structure of the characteristics of interest can help in choosing among alternative estimation strategies possessing both good model- and design-based properties. It may even help assess whether the sample is large enough for purely design-based inference.

One example of an estimation strategy with good model- and design-based properties is estimating T with t_{ratio} based on a without-replacement simple random sample. This strategy is nearly design unbiased as well as unbiased under the ratio model. An unbiased estimator for its model variance, $E_M\left[(t_{ratio} - T)^2\right]$, when $\sigma_k^2 \propto x_k$ is

$$v_{ratio} = \left[\left(\frac{N\bar{x}_U}{n\bar{x}_S}\right)^2 - \left(\frac{N\bar{x}_U}{n\bar{x}_S}\right)\right] \sum\nolimits_{k \in S}\left\{\left(y_k - \frac{\bar{y}_S}{\bar{x}_S}x_k\right)^2 \Big/ - \left(1 - \frac{x_k}{n\bar{x}_S}\right)\right\},$$

which is also a nearly unbiased estimator for the strategy's design variance under mild conditions. Empirical studies have shown that variance estimators with both good model- and design-based properties tend to produce confidence intervals with closer-to-predicted coverage rates than purely model- or design-based ones.

Phillip S. Kott

See also Convenience Sampling; Cutoff Sampling; Design-Based Estimation; Finite Population; Ratio Estimator; Simple Random Sample

Further Readings

Kott, P. S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, *129*, 263–277.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Valliant, R., Dorfman, R. M., & Royall, R. M. (2000). *Finite population sampling and inference*. New York: Wiley.

Mode of Data Collection

Within the context of survey operations, *mode* refers to an employed method or approach used for the collection of data. For example, surveys may be conducted face to face, by telephone, mail, or Internet (the four most commonly used modes), or through other types of approaches (such as interactive voice response [IVR], disk-by-mail, etc.) or combinations of modes. Modes can differ along a number of dimensions, including whether an interviewer is present, how the questions are presented and the responses recorded, the infrastructure required, field time, and costs.

One of the primary distinctions between modes of data collection is the presence or absence of an interviewer. When an interviewer is present, the survey questions are generally read to the respondent, and the mode is referred to as *interviewer-administered*. Telephone and face-to-face (in-person) surveys are examples of interviewer-administered data collection. When an interviewer is not present and the respondent must deal directly with a paper or electronic questionnaire, the mode is generally said to be *self-administered*. Examples of these include mail and Internet-based surveys.

The method of presenting the questions and receiving the responses also defines the mode of data collection. Questions presented visually are typically read by respondents, whereas those presented verbally are heard by respondents. The way in which the respondent receives the stimuli of the question has been shown to affect how a person responds to a particular survey question. Likewise, responses provided to survey questions can be written by hand, typed, or spoken. Each of these methods presents different memory and perception issues. Questions and response options that are read *to* respondents generally need to be shorter than those that are read *by* respondents, because of working memory limitations. When response categories are received visually, respondents tend to choose categories early in the list (a *primacy effect*). When they are received aurally, respondents tend to choose categories toward the end of the list (a *recency effect*). Thus, researchers must pay special attention to possible mode effects on data quality, especially in mixed-mode surveys in which some answers to a question come from respondents who were contacted via one mode (e.g., mail) and other answers to these same questions come from a different mode (e.g., telephone).

The infrastructure (and thus the financing) needed to conduct a survey also differs by mode. A self-administered, Web-based survey of several thousand individuals could potentially be carried out by an individual person, while a face-to-face survey of the same size would require a staff of interviewers and field managers. If a telephone survey is being considered, a centralized telephone interviewing facility often is required. Within any specific mode, the infrastructure requirements may depend on the sample size and on the needs of the researcher. A telephone survey of a few hundred individuals could be conducted by a team of students using paper-and-pencil questionnaires. However, a national face-to-face survey of several thousand will probably require a large survey research center with appropriate staffing and experience. Some recommend that the same professional and scientific standards be followed regardless of the particular structure of a research project. While a team of students can conduct the interviews, they should be trained on ethical issues of conducting research and on appropriate interviewer behavior (e.g., how to read questions) at the same level as staff in a professional research organization. This will include very specific instructions about whether questions are to be read as worded, whether and how the interviewers should probe respondents, and whether clarification or elaboration can be given by the interviewer.

The time it takes to complete data collection depends on how long it takes to contact respondents, administer questionnaires, and return the data for processing; and this will vary by mode. Data collection modes that are centralized (such as a telephone facility or Web-based data collection system) can typically collect data in a relatively short amount of time. Some surveys by telephone and Internet are conducted overnight. In contrast, mail surveys must take into account how long it takes for the questionnaire package to reach the respondent, the time required for the person to complete the questionnaire, and then the return mailing time. This can take up to 4 to 6 weeks, and even longer when follow-up mailings are used.

Data collection modes also differ in terms of cost. Variations in cost are dependent on the amount of effort, resources, and infrastructure required to collect the data, as well as the sheer size of the effort in terms of numbers of respondents. Costs can be divided into fixed and variable categories. Fixed costs are those that would be required even if only one respondent were sampled. They do not vary with the sample size. Variable cost will go up (although not necessarily linearly) with increases in sample size. For example, the costs of *designing* a mail survey will be identical whether 1 or 1,000 questionnaires are to be mailed. However, the costs for printing, mailing, and data entry will vary depending on the number of questionnaires mailed and returned. In contrast, Web surveys are good examples of surveys with potentially high fixed costs (depending on whether the survey firm needs to purchase and maintain computer hardware), but with low per-unit variable costs (i.e., the cost of fielding an extra 1,000 cases is low because using the Internet is essentially free data that do not need to be entered by survey staff).

Sometimes considerations of survey coverage and nonresponse make it necessary to combine data collection modes. For example, to obtain a required response rate, a researcher may first need to mail the questionnaire to sample units, then conduct a telephone and/or in-person survey for nonrespondents to the mail survey. Also, within a single mode of administration, facets of different modes can be combined, as in when show cards are used to present response categories visually in a face-to-face survey, rather than simply having the response options read to the respondent. The future is likely to involve more complex combinations of data collection modes as researchers seek to reduce costs, maintain response rates, and take advantage of technological advances.

Matthew Jans

See also Disk by Mail; Face-to-Face Interviewing; Field Survey; Interactive Voice Response (IVR); Internet Survey; Mail Survey; Mixed-Mode; Mode; Mode Effects; Mode-Related Error; Multi-Mode Surveys; Primacy Effect; Recency Effect; Self-Administered Survey; Survey Costs; Telephone Survey; Web Survey

Further Readings

de Leeuw, E. D. (1992). *Data quality in mail, telephone, and face to face surveys*. Amsterdam: T. T. Publikaties.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design* (2nd ed.). New York: Wiley.

Fowler, F. J., Jr. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Mode-Related Error

Face-to-face (in-person) surveys, telephone, mail, and Web surveys are common types of data collection modes in current survey research. These modes can be classified into two categories: (1) *self-administered* versus *interviewer-administered*, depending on whether interviewers are involved in interviewing, (2) and *paper-and-pencil* versus *computer-assisted*, depending on whether computerized instruments are employed in surveys. Currently, the two types of classification for survey modes are widely used for comparing response rates or survey errors. Total survey errors—sampling, coverage, nonresponse, measurement, and processing error—do not equally emerge from each mode. That is, the type and magnitude of error varies with the mode.

Sampling error is not directly influenced by modes. All sample surveys inevitably encounter this error, due to the fact that the whole target population is not selected as a sample. However, its magnitude varies, depending mostly on sample size. Generally, sampling error decreases as sample size increases, regardless of mode.

Coverage error arises from the mismatch between a sample frame and a target population frame. It is not dependent on the mode itself, but on the quality of the sample frame that the mode employs. For instance, in 2000 in the United States, the penetration rate of households with a landline telephone was more than 95% (including cell phone only households), whereas household Internet access was approximately 50%. This means that a Web mode is much more susceptible to coverage error than a telephone method in household surveys. Among data collection modes, the Web method seriously suffers from this error because of a poor frame, a disparity in Internet penetration between the poor and rich, and the existence of users with multiple email accounts.

For nonresponse error, nonresponse rates can affect the amount of error. However, it is important to note that reduction in the nonresponse rate does not always affect nonresponse error. There are two types of nonresponse in surveys: item and unit nonresponse. The former occurs when respondents avoid reporting one or more specific questions, whereas the latter arises primarily from noncontact, refusal, and inability to answer (e.g., a language barrier). Item nonresponse diminishes in interviewer-administered modes, whereas the occurrence of unit nonresponse varies across modes. In general, unit nonresponse rates for face-to-face surveys are the lowest, followed by higher nonresponse rates for telephone and mail surveys, with Web surveys having the highest unit nonresponse. Generally, interviewer-administered modes have higher response rates than self-administered.

There are a number of sources of measurement error, including social desirability effects, interviewers, respondents, questionnaires, and so forth. Social desirability can be a large threat to survey validity. Respondents tend to provide socially desirable and avoid socially undesirable responses in surveys. Indeed, interviewer-administered modes are more susceptible to social desirability bias than self-administered modes because respondents are reluctant to disclose socially stigmatized behaviors in the presence of interviewers, especially when queried about sensitive topics. Also, social desirability biases are reduced in computer-assisted self-interviewing (CASI) and audio computer-assisted self-interviewing (ACASI). CASI and ACASI lead to less item nonresponse for sensitive questions—the number of sex partners, abortion, drug or substance use, and so on—than computer-assisted personal interviewing (CAPI), computer-assisted telephone interviewing (CATI), and self-administered paper-and-pencil interviewing (PAPI). Directive probing or interviewers' characteristics (i.e., gender and race) may also lead to interviewer effects that vary responses across interviewers in interviewer-administered modes.

The questionnaire itself also can affect respondents. Context effects occur when respondents are affected by previous questions or their prior responses when they answer a subsequent question. Compared to self-administered modes, interviewer-administered modes produce more of such errors. Besides question order, the order in which response options are presented also can affect respondents (i.e., response order effects). Respondents have a tendency to choose the first response option when options are presented visually (i.e., primacy effects). This usually happens in self-administered modes and interviewer-administered modes with show cards. On the other hand, respondents are likely to choose the last option when they listen to response options in interviewer-administered surveys (i.e., recency effects).

Thus, there is ample evidence that the choice of the mode for data collection is a very important one for the survey researcher to make. Furthermore, researchers must pay especially close attention to possible mode-related differences, including differential types of errors, when conducting mixed-mode surveys.

Geon Lee

See also Audio Computer-Assisted Self-Interviewing (ACASI); Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Coverage Error; In-Person Surveys; Mail Surveys; Measurement Error; Mixed-Mode; Mode Effects;

Mode of Data Collection; Nonresponse Error; Paper-and-Pencil Interviewing (PAPI); Primacy Effects; Recency Effects; Sampling Error; Sensitive Topics; Social Desirability; Telephone Surveys; Total Survey Error (TSE); Web Survey

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Multi-Level Integrated Database Approach (MIDA)

The multi-level integrated database approach (MIDA) is an enhancement to survey sampling that uses databases to collect as much information as practical about the target sample at both the case level and at various aggregate levels during the initial sampling stage. The goal of MIDA is to raise the final quality, and thus the accuracy, of survey data; it can do this in a variety of ways.

Building an MIDA

The following description of MIDA uses the example of national samples of U.S. households based on addresses and as such is directly appropriate for postal and in-person samples. However, similar approaches can be applied to other modes and populations (e.g., national random-digit dialing [RDD] telephone samples, panel studies, list-based samples, and local surveys).

The first step in MIDA is to extract all relevant public information at both the case level and aggregate levels from the sampling frame from which the sample addresses are drawn. In the United States, general population samples of addresses are typically nearly void of household-level information. However, U.S. address samples are rich in aggregate-level information. Address or location, of course, is the one known attribute of all cases, whether respondents or nonrespondents. Moreover, address-based sampling frames are typically based on the U.S. Census and as such the appropriate census data from blocks, tracts, place, and so on are part of the sampling frame and are linked to each address.

The second step is to augment the sampling frame by linking all cases in the sample to other databases. At the case level, that means linking the addresses to such sources as telephone directories, credit records, property records, voter registration lists, and many other public sources. The information obtained includes whether a match was or was not found (e.g., listed in telephone directory or not), and, if matched, whatever particular information is available (e.g., names, telephone numbers, credit reports, voter registration status).

At the aggregate level, this means merging information from sources other than those in the sampling frame. Examples of aggregate-level data beyond that from the census that could be appended are consumer information from such sources as Claritas's *PRIZM NE* and Donnelley Marketing's *FIND Index*, voting information from national elections, and data on such other matters as vital statistics, crime rates, religion, public housing, HIV/STD rates, and public welfare utilization.

The linked data include information from multiple levels of aggregation. The multi-level analysis starts with household-based data and includes neighborhood-level data from census tract and zip code-based data sources, community-level data from the census, election counts, crime rates, and other sources, and higher-level aggregations (e.g., metropolitan areas and census divisions).

The third step is to take information gained from the initial case-level linkages to secure additional information. For example, securing a name and telephone number from a telephone directory search can lead to households being found in databases when a mere address was insufficient to allow a match. Also, once a respondent is identified, links to that person in addition to household-level matching can be carried out. Thus, the process of augmenting the sampling frame is iterative and continues during the data collection phase.

The final step is to record, process, clean, and maintain a large amount of paradata for each case. This includes having interviewers systematically record information about the sample residence (e.g., dwelling type, condition of dwelling), contacts or call attempts, interactions with household members (including contacts that end as refusals), and observations on the composition and demographics of the household.

Using and Benefiting From an MIDA

The multi-level information in this greatly enriched sampling frame can be used to advantage for data collection, nonresponse measurement and adjustment, interview validation, and substantive analysis.

First, more information on the target sample makes data collection both more efficient and more effective. This information can be used both to assist making contact with the household and to help tailor approaches once contact is made.

Second, the information in the MIDA-augmented sampling frame can be used to measure and adjust for nonresponse error. Having a wide range of case-level and aggregate-level information is important for testing the representativeness of the achieved sample across as many variables as possible and because surveys covering different topics are likely to have different nonresponse profiles (e.g., nonvoters underrepresented in political surveys; the wealthy in the Survey of Consumer Finance). Having more relevant information on nonrespondents allows for better modeling of possible nonresponse bias and the creation of weights that more fully account for the biases and also has the particular advantage of having augmented data for all sample cases.

Third, MIDA can facilitate interview validation procedures by allowing the information from the databases to be used along with recontacts to help corroborate that interviews were truly and correctly done.

Finally, for respondents, the case-level and aggregate-level data in the augmented sampling frame can be utilized for crucial substantive analysis. While most case-level information would come from the interviews with the respondents, the added case-level data would include both information uncovered in any particular survey and data that can be used to corroborate information reported by respondents. Additionally, aggregate-level information is of great utility. Research has demonstrated that contextual aggregate-level geographic effects in general, and neighborhood characteristics in particular, influence a wide range of attitudes and behaviors independent of the attributes of individuals. The coding of a rich array of aggregate-level data from the sampling frame and a wide range of databases facilitates such contextual analysis and makes it a regular part of survey analysis rather than an occasional approach carried out only when special multi-level data are added, often after the fact, to standard surveys. In sum, the information in the augmented sampling frame that can be used to assist data collection and adjust for nonresponse bias also can be used for multi-level contextual analysis.

Tom W. Smith

See also Address-Based Sampling; Nonresponse Error; Paradata; Refusal Report Form (RRF); Sampling Frame; Validation

Further Readings

Chaves, M. (2007). *The General Social Survey: Innovation and dissemination comments for NSF workshop on the GSS*. Retrieved April 14, 2008, from http://www.nsf.gov/pubs/2007/nsf0748/nsf0748_9.pdf

Smith, T. W. (2007, November). *Using the multi-level integrated database approach*. Paper presented to the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

MULTI-MODE SURVEYS

Multi-mode surveys (sometimes called "mixed-mode surveys") involve collecting information from survey respondents using two or more modes and combining the responses for analysis. Multi-mode surveys have become increasing popular because of the rise of new modes of data collection, the impact of computer technology, and decreasing response rates to traditional survey modes (particularly telephone surveys). The development of new modes of data collection has expanded the methods available to survey researchers. Multi-mode survey designs are extremely flexible when various combinations of modes can be employed to adapt to the particular needs of each research study. Multi-mode surveys are often used to compensate for coverage biases of individual modes and to increase overall response rates. However, these reductions in coverage and nonresponse must be balanced with potential increases in measurement error that may arise from combining responses collected using different modes.

Survey designs involve choosing the optimal mode or combination of modes while minimizing overall total survey error (coverage, sampling, nonresponse, and measurement). The decision of whether to use multiple modes for data collection involves several issues. Surveyors should consider the best mode or modes for the population and research of interest; some populations may not have access to a particular mode or may prefer to be surveyed by one mode, contact information may only be available for one mode of communication, and some questions or topics may lend themselves to a particular mode. Survey modes can be chosen to increase coverage of the population of interest (e.g., dual-sampling frame designs) and to minimize nonresponse bias resulting from differences between respondents and nonrespondents. Mode decisions are almost always influenced by the budget constraints of the particular study; often less expensive modes are used before more expensive modes to reduce overall data collection costs. Last, choices about survey mode are also guided by how quickly the data needs to be collected and whether the survey must be fielded within a particular time period.

Types of Multi-Mode Surveying

There are four general types of multi-mode survey designs that can involve various combinations of modes of data collection.

Sampling Via One Mode, Data Collection Via Another

First, the most common type of multi-mode survey occurs when one mode is used to collect data from some members of the sample and one or more additional modes are used to collect data from other sample members. This type of multi-mode survey design can involve concurrent or sequential data collection. Multiple modes can be employed to collect data at one time (e.g., a paper survey with a Web option) or over a period of time (e.g., respondents are mailed a questionnaire and then nonrespondents are later surveyed by telephone or personal interview).

Using More Than One Mode to Gather Data From the Same Respondent

Another type of multi-mode survey design uses a different mode to collect certain types of information from the same respondent. For example, personal interview respondents may answer sensitive questions using computer-assisted self-interviewing (CASI) or may be asked to complete a consumer diary on paper and return it by mail.

Changing Modes Over Time in Longitudinal Studies

A third type of multi-mode design involves surveying members of the same sample or of different samples using multiple modes over time, where the survey mode changes for different periods or phases of data collection. For example, face-to-face personal interviews may be used for the initial period of data collection in a longitudinal survey, but subsequent data collection periods may survey respondents by telephone, mail, or Internet.

Combining Data From Different Modes in the Same Larger Survey

The final type of multi-mode survey involves combining independently collected data from different samples, subgroups, or populations. For example, many international surveys are conducted in which data may be collected in one country by personal interviews and in another country by telephone or mail and then combined for analysis. In addition, data may also be collected independently in different studies and then combined for comparative analysis (e.g., comparing data collected for a particular city or state to nationally collected data).

Data Quality

Combining data collected from different survey modes for analysis may introduce mode-related measurement error and reduce data quality. Mode effects arise because social, cultural, and technological factors associated with particular modes influence how respondents complete the survey response process. Respondents' answers to survey questions are influenced by how information is communicated with respondents, their varying familiarity with and use of the medium or technology, whether the respondent or an interviewer controls the delivery of the survey questions, and the presence of an interviewer. To reduce measurement differences, optimal design of survey questionnaires for multi-mode surveys should focus on presenting an equivalent stimulus to respondents across different modes. This type of unified or universal mode design should recognize how differences in meaning may depend on how information is communicated with respondents. In addition, questionnaire design for multi-mode surveys in which most of the responses are expected by one mode should design for the primary mode and allow it to inform the design of the questionnaires for the secondary or supplementary modes.

Leah Melani Christian and Kelly N. Foster

See also Aural Communication; Mixed-Mode; Mode; Mode Effects; Mode of Data Collection; Mode-Related Error; Survey Costs; Visual Communication

Further Readings

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality*. New York: Wiley.

de Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*, 233–255.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method: 2007 update with new Internet, visual, and mixed mode guide*. Hoboken, NJ: Wiley.

Multiple-Frame Sampling

Most survey samples are selected from a single sampling frame that presumably covers all of the units in the target population. *Multiple-frame sampling* refers to surveys in which two or more frames are used and independent samples are respectively taken from each of the frames. Inferences about the target population are based on the combined sample data. The method is referred to as *dual-frame sampling* when the survey uses two frames.

Sampling designs are often dictated by several key factors, including the target population and parameters of interest, the population frame or frames for sampling selection of units, the mode of data collection, inference tools available for analyzing data under the chosen design, and the total cost. There are two major motivations behind the use of multiple-frame sampling method: (1) to achieve a desired level of precision with reduced cost and (2) to have a better coverage of the target population and hence to reduce possible biases due to coverage errors. Even if a complete frame, such as a household address list, is available, it is often more cost-effective to take a sample of reduced size from the complete frame and supplement the sample by additional data taken from other frames, such as telephone directories or institutional lists that might be incomplete but less expensive to sample from. For surveys of human populations in which the goal is to study special characteristics of individuals, such as persons with certain rare diseases, a sample taken from the frame for general population health surveys is usually not very informative. Other frames, such as lists of general hospitals and/or special treatment centers, often provide more informed data as well as extended coverage of the target population.

There are, however, unique features, issues, and problems with inferences under multiple-frame sampling, which require unique treatments and special techniques. Let $Y = \sum_{i=1}^{N} y_i$ be the population total of a study variable y, where N is the overall population size. Suppose there are three frames: A, B, and C. Each of them may be incomplete, but together they cover the entire target population. Let s_A, s_B, and s_C be the three independent samples taken respectively from frames A, B, and C. The basic question is how to estimate Y using all three samples. It turns out that none of the samples can directly be used if the frames are incomplete. The most general picture is that the three frames divide the target population into seven disjoint domains: A, AB, ABC, AC, B, C, and BC, where A contains population units from frame A but not covered by B or C, AB includes all units from both A and B but not C, ABC represents the set of units covered by all three frames, and so on. If, for instance, frames B and C are nonoverlapping, then the domain BC vanishes. We can rewrite the overall population total as $Y = Y_A + Y_B + Y_C + Y_{AB} + Y_{AC} + Y_{BC} + Y_{ABC}$, where, for instance, Y_A is the population total for domain A. Each of the three samples can also be partitioned according to the involved population domains: $s_A = s_a \cup s_{ab} \cup s_{ac} \cup s_{abc}$, $s_B = s_b \cup s_{ba} \cup s_{bc} \cup s_{bac}$ and $s_C = s_c \cup s_{ca} \cup s_{cb} \cup s_{cab}$, where, for instance, units in both s_{ab} and s_{ba} are selected from the domain AB, s_{ab} is from frame A, whereas s_{ba} is from frame B, indicated by the first letter in the subscript. Estimation of Y is typically carried out through the estimation of domain totals using relevant sample data.

Major issues and problems with estimation under multiple-frame sampling include but are not restricted to the following:

1. *Frame membership identification for all sampled units*. This is required in order to post-stratify samples from different frames into appropriate population domains. Additional questions regarding frame memberships need to be included for data collection.

2. *Estimation of domain totals using multiple samples*. For instance, both s_{ab} (sampled from Frame A but also found in Frame B) and s_{ba} (sampled from Frame B but also found in Frame A) are selected from the domain AB and need to be combined to estimate Y_{AB}. This may not be as straightforward as it appears to be, since the sampling designs used for frames A and B could be different. Obtaining efficient domain estimates can be challenging, especially for domains such as ABC. where all three samples s_{abc}, s_{bac}, and s_{cab} need to be combined.
3. *Lack of information on the domain population sizes*. Under certain designs one may, for instance, have an estimator readily available for the domain mean $\overline{Y}_{AB}$ and estimation of the total Y_{AB} requires that the domain size N_{AB} be either known or easily estimable, which is not always the case.
4. *Identifying and removing duplicated units from multiple-frame samples*. This is required by some methods based on pooled samples when no single unit is allowed to be used more than once.
5. *Handling the extra variation induced by the random sample sizes*. Even if all the initial sample sizes are fixed, the sizes of the post-stratified samples are still random. This creates difficulties for variance estimation.
6. *Use of auxiliary information for estimation*. The known auxiliary population information could be for the entire target population or for specific frames or for both. Incorporating such information with multiple-frame samples requires approaches that differ from the conventional single-frame methodologies.

When all frames are complete, multiple-frame sampling becomes the so-called multiple surveys in which several independent samples are taken from the same target population.

Changbao Wu

See also Auxiliary Variable; Coverage Error; Dual-Frame Sampling; Frame; Mode of Data Collection; Post-Stratification; Survey Costs; Target Population

Further Readings

Hartley, H. O. (1974). Multiple-frame methodology and selected applications. *Sankhya, Series C*, *36*, 99–118.

Lohr, S., & Rao, J. N. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, *101*, 1019–1030.

Skinner, C. J., Holmes, D. J., & Holt, D. (1994). Multiple-frame sampling for multivariate stratification. *International Statistical Review*, *62*, 333–347.

Multiple Imputation

Multiple imputation (MI) is actually somewhat of a misnomer. The phrase is best understood as the name for a *post-imputation variance estimation tool* that involves repetitions of the imputation process. The father of multiple imputation, Donald Rubin, originally envisioned MI as a tool for the preparation of public use files (PUFs). He advocated that data publishers use MI in order to simplify and improve the analyses conducted by PUF consumers. So far, few data publishers have adopted MI. More usage of MI has been found in highly multivariate analyses with complex missing data structures, such as in the scoring of standardized tests with adaptive item sampling. In that literature, the multiple imputations are most often referred to as *plausible values*.

Motivation

MI is most commonly used in conjunction with Bayesian imputation methods, in which samples drawn from the posterior distribution of the missing data given the observed data are used to fill in the missing values. However, as long as there is some element of randomness in the imputation process, one can imagine executing the process multiple times and storing the answers from each application (i.e., replication). The variance of a statistic of interest across these replications can then be calculated. This variance can be added to the "naïve" estimate of variance (obtained by treating all imputed data as if they were observed) to produce a variance estimate for the statistic that reflects the uncertainty due to both sampling and imputation. That is the essence of multiple imputation.

Controversy

There is a long-standing heated debate within the community of survey research statisticians about the utility of MI for analyses unanticipated by the data publisher. It is easy to find examples where mechanical application of MI results in over- or undercorrection. Rubin has a theorem that identifies the class of imputation procedures that can be used in conjunction with MI to obtain asymptotically valid inferences for a given statistic. He labels such imputation procedures as "proper." However, a series of debates in the 1990s, culminating in a trio of 1996 papers, demonstrated that proper imputation methods are difficult to construct. Moreover, an imputation procedure that is proper for one analysis might be improper for another analysis.

Basic Formulae

Suppose that the entire imputation process of choice is repeated m times and that all m imputed values are stored along with the reported data. Conceptually, the process produces m completed data sets representing m replicates of this process. If there were originally p columns with missing data, then there will be mp corresponding columns in the new multiply imputed dataset. The user then applies his or her full-sample analysis procedure of choice m times, once to each set of p columns. Suppose that $\hat{\theta}_{Ik}$ is the point estimate of some parameter, θ, based on the kth set of p columns. (The subscript, $_I$, indicates employment of imputed data.) Also suppose that $\hat{V}_{Ik}$ is the variance estimate for $\hat{\theta}_{Ik}$ provided by the standard complex survey analysis software when applied to the kth set of p columns.

Assuming that the imputation method of choice has a stochastic component, such as imputation that is based on a linear regression model to predict imputed values from covariates, multiple imputations can be used to improve the point estimate and provide better leverage for variance estimation. Rubin's point estimate is $\hat{\theta}_m = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_{Ik}$, and his variance estimate is

$$T_m = \frac{1}{m}\sum_{k=1}^{m}\hat{V}_{Ik} + \frac{m+1}{m}\frac{1}{m-1}\sum_{k=1}^{m}\left(\hat{\theta}_{Ik} - \hat{\theta}_m\right)^2 = \bar{U}_m + B_m.$$

With a proper imputation method, $\bar{U}_\infty = \lim_{m\to\infty}\bar{U}_m$ closely approximates the variance of an estimate of θ that could be produced if all sample members responded, and $B_\infty = \lim_{m\to\infty} B_m$ approximates the variance caused by both the missing data and the imputation procedure.

Pathological Examples

Consider now what can go wrong from the application of multiple imputation to an improper imputation procedure. If the imputation procedure is a deterministic method (i.e., has no stochastic component), such as mean imputation or nearest-neighbor imputation, then $B_m = 0$ (i.e., no variability in estimates across the imputation replicates), leading to an underestimate of $var(\hat{\theta}_m)$.

Overestimation of variances is possible as well, as in the famous example of Robert Fay. Here a more peculiar but less subtle example is considered. Suppose that for the variable Y, the data publisher randomly picks one respondent and imputes that single value to all nonrespondents. Suppose further that there are two domains, A and B, and that the parameter of interest is the difference in the mean of Y across them, despite the fact that, unbeknown to the analyst, this difference is zero. Assume a simple random sample with replacement of size n with "missingness" completely at random. Assume that the response rates in the two strata, R_A and R_B, are unequal.

Then $\hat{\theta}_{Ik} = \bar{y}_{AR}R_A - \bar{y}_{BR}R_B + Y_{Rk}(R_B - R_A)$, where $\bar{y}_{AR}$ and $\bar{y}_{BR}$ are the means among respondents in the two domains, and Y_{Rk} is the universal donor chosen on multiple impute k. From this, $\hat{\theta}_\infty = \lim_{m\to\infty}\hat{\theta}_m = \bar{y}_{AR}R_A - \bar{y}_{BR}R_B + \bar{y}_R(R_B - R_A)$, where $\bar{y}_R$ is the overall respondent mean. Note that $var(\hat{\theta}_\infty)$ is inversely proportional to the sample size. However, $\left(\hat{\theta}_{Ik} - \hat{\theta}_\infty\right)^2 = (R_B - R_A)^2(Y_{Rk} - \bar{y})^2$, so $B_\infty = (R_B - R_A)^2\sigma^2$, where σ_2 is the element variance of Y. Clearly this B_∞ does not decrease with the sample size. There is also a term in $\hat{V}_{Ik}$ and therefore in $\bar{U}_\infty$ that does not decrease with the sample size. Thus, T_∞ is too large by an order of magnitude. The pathology here is caused by the fact that the publisher ignored the domains of interest to the consumer.

Guidelines

From the pathological examples, we see that it is possible through the choice of imputation method to induce either too much or too little variability among the plausible values. What method will induce just the right amount? How to choose m? These are open questions for all but the simplest analyses. Bayesian Markov chain Monte Carlo (MCMC) methods are probably a good choice, but the search for simpler alternatives continues.

There is no general theory on how to optimally choose m. Although $var(\hat{\theta}_m)$, $var(\bar{U}_m)$, and $var(B_m)$ are all nonincreasing functions of m, the computing demand is an increasing function of m, and examples have been discovered in which bias (T_m) also increases with m. A common choice is to use $m = 5$. A larger number may be particularly desirable if the item nonresponse rate is high.

Despite the discomfort caused by the lack of firm answers to these questions, no better post-imputation variance estimation methods have been found that apply to multivariate analyses, such as a regression of one variable on two or more other variables, each with a distinct missing data pattern. The alternatives that have been identified are mostly applicable only to univariate statistics, with some extensions to multivariate analyses in which variables are missing in tandem or block style instead of Swiss cheese style. However, if the publisher did not condition the imputation in such a way as to protect the relationships of interest to the user, then the user may wish to consider replacing the published set of plausible values with his or her own.

David Ross Judkins

See also Hot-Deck Imputation; Imputation; Variance Estimation

Further Readings

Christian, P. R., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). New York: Springer.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, *91*(434), 490–498.

Graham, J. W., Olchoski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213.

Harel, O., & Zhou, X. H. (2007). Multiple imputation: Review of theory, implementation, and software. *Statistics in Medicine*, *26*, 3057–3077.

Kim, J. K., Brick, J. M., Fuller, W. A., & Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society B*, *68*(3), 509–521.

Rao, J. N. K. (1996). Variance estimation with imputed survey data (with discussion). *Journal of the American Statistical Association*, *91*(434), 499–520.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Schafer, J. L. (1996). *Analysis of incomplete survey data*. New York: Chapman & Hall.

Multiplicity Sampling

Multiplicity sampling is a probability sampling technique that is used to enhance an existing sampling frame by adding elements through a form of network sampling. It is especially useful when surveying for rare attributes (e.g., rare hereditary diseases).

In sample surveys, population elements are linked to units in the sampling frame, or frame units. For example, persons can be linked to a specific household by familial relationship among persons residing in the same housing unit. A *counting rule* identifies the linkage between population elements and frame units. In most surveys, population elements are linked to one and only one frame unit, and thus there is a one-to-one correspondence of element to unit. For multiplicity sampling, a counting rule is established that defines the linkage between population elements and frame units in which one or more population elements are linked to one or more frame units. The counting rule defines a one-to-many or a many-to-many correspondence between population elements and frame units. The count of frame units linked by the counting rule to each population element is called the *multiplicity* of a frame unit.

For an unbiased estimate of the number of population elements, the design-based sampling weight for each selected frame unit is adjusted for the number of frame units linked to the population element by dividing the sampling weights by the multiplicity. The multiplicity is needed for only those units selected in the sample. Multiplicity sampling uses the linkage of the same population element to two or more frame units to allow the sample of frame units to identify more population elements.

For example, in a household survey to estimate the frequency of a target condition in a population, the standard household survey would enumerate only persons with the target condition in sampled households. With multiplicity sampling, a counting rule based on adult biological siblings residing in households would identify a person with a specific attribute (the population element) linked to their own household and to the households of his or her adult biological siblings. Each sampled household member would be asked, (a) if you or an adult biological sibling have the specific condition, (b) the number of adult siblings with the condition, and (c) the number of households containing adult biological siblings. The person with the attribute would be identified with all of these households, not only their own household. Each frame unit in the sample would be assigned the count of adult siblings with the condition, and the multiplicity would be the number of households containing adult biological siblings of the person with the condition. The multiplicity-adjusted sampling weight is the design-based sampling weight for the household member divided by the multiplicity.

The sampling variance would be computed using the standard variance estimator appropriate for the sampling design. Because the multiplicity for each sampled frame unit will vary, the multiplicity-adjusted sampling weights often exhibit more variation than the design-based sampling weights before the multiplicity adjustment and can be expected to increase the sampling variance relative to the sampling.

Multiplicity sampling is an option when population elements with the target condition are rare and the costs of the large sample to identify an adequate number of population elements are beyond the survey resources. Multiplicity sampling requires a clear workable counting rule that can achieve an accurate count of the multiplicity for each sampling unit.

Frank Potter

See also Elements; Multiple-Frame Sampling; Network Sampling; Probability Sample; Rare Populations; Sampling Frame; Snowball Sampling; Unit

Further Readings

Czaja, R. F., Snowdon, C. B., & Casady, R. J. (1986). Reporting bias and sampling errors in a survey of a rare population using multiplicity counting rules. *Journal of the American Statistical Association*, *81*, 411–419.

Sirken, M. G. (1972). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, *67*, 224–227.

Sirken, M. G., & Levy, P. S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *Journal of the American Statistical Association*, *69*, 68–73.

MULTI-STAGE SAMPLE

A multi-stage sample is one in which sampling is done sequentially across two or more hierarchical levels, such as first at the county level, second at the census track level, third at the block level, fourth at the household level, and ultimately at the within-household level.

Many probability sampling methods can be classified as single-stage sampling versus multi-stage sampling. Single-stage samples include simple random sampling, systematic random sampling, and stratified random sampling. In single-stage samples, the elements in the target population are assembled into a sampling frame; one of these techniques is used to directly select a sample of elements. In contrast, in multi-stage sampling, the sample is selected in stages, often taking into account the hierarchical (nested) structure of the population. The target population of elements is divided into first-stage units, often referred to as *primary sampling units* (PSUs), which are the ones sampled first. The selected first-stage sampling units are then divided into smaller second-stage sampling units, often referred to as *secondary sampling units* (SSUs), which are sampled second. This process continues until the actual elements, also referred to as the *ultimate sampling units*, are reached.

For example, to obtain a national sample of elementary public school students, one can divide the target population of students into elementary schools in the United States, which are used as first-stage sampling units (i.e., the PSUs). Sample schools are selected at the first stage of sampling. A sampling frame (list) of students is then assembled for each selected school. At the second stage of sampling, a sample of students is selected from each selected school. This design is a two-stage sample.

In another example, to obtain a national sample of housing units, one can divide the target population of housing units into counties, which are used as the first-stage sampling units (i.e., the PSUs). A sample of counties is then selected. Within each selected county, the target population of housing units is divided into census tracts. A sample of census tracts is drawn from within each selected county. The census tracts would be considered the SSUs. Within each selected census tract, the target population is divided into census blocks. A sample of census blocks is drawn from each selected census tract. The census blocks would be considered the third-stage sampling units. Within each selected census block, a sampling frame (list) of all housing units is assembled. A sample of housing units is then sampled from each of the selected census blocks. The housing units would be considered the fourth-stage sampling units. This design is a four-stage sample.

In both examples, the hierarchical structure of each population was used. Also note that there is a size ordering in the second example—there are more census blocks in the United States than there are census tracts, and there are more census tracts than counties. One must use an appropriate method of selection at each stage of sampling: simple random sampling, systematic random sampling, unequal probability sampling, or probability proportional to size sampling. Also, one can incorporate stratified sampling procedures to select a stratified multi-stage sample. In the previous examples, one would at a minimum want to stratify the first-stage sampling units, elementary schools and counties, by the four census regions.

Multi-stage sampling is widely used for several reasons. First, a sampling frame of the elements may not exist or may be too expensive to construct. In the two examples given, no complete list of all elementary public school students in the United States exists, and no complete list of all housing units in the United States exists. It is therefore not possible to draw a single-stage sample of these elements. In this situation, one must take advantage of the hierarchical structure of the population and design a multi-stage sample. Second, even if a sampling frame of the elements exists, it may be more cost-effective to use a multi-stage sample design. For example, in a national in-person interview survey, the cost of travel to a widely dispersed sample of housing units would lead to a very high cost of data collection. In a multi-stage design, the interviewers travel to the selected census block, where they attempt to contact all of the sample housing units in that census block. If two census blocks are selected from each census tract, then there is another census block in the same census tract that contains sampled housing units. Thus, the cost of travel associated with each sample housing units in the multi-stage design is much lower than in a single-stage design.

For a fixed sample size of elements, a multi-stage sample design is almost always less efficient than a simple random sample. The design of a multi-stage sample does, however, allow for some control of the loss of efficiency. For example, in the previous two-stage sample design example, one can sample more schools and select fewer students per school to reduce the loss in efficiency compared to a simple random sample of students. The design effect (*deff*) is the most commonly used statistic to measure the loss in efficiency from using a two-stage or a multi-stage sample design. One also needs to be aware that the usual formulas for standard errors under simple random sampling do not apply. Variance estimation methods for complex sample design must be used to obtain correct standard errors.

Michael P. Battaglia

See also Design Effect (*deff*); Elements; Primary Sampling Unit (PSU); Probability Proportional to Size (PPS) Sampling; Segments; Simple Random Sample; Stratified Sampling; Survey Costs; Systematic Sampling; Variance Estimation

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory* (Vol. I: *Concepts and discussion*). New York: Wiley.

Kish, L. (1965). *Survey sampling* (2nd ed.). New York: Wiley.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.

Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Mutually Exclusive

Response options to a survey question are *mutually exclusive* when only one response option can be true for a single respondent. Consider a survey question that asks respondents, *How long do you spend commuting each day (round trip): less than 15 minutes, 15 to 30 minutes, 30 minutes to one hour, or one hour or longer?* A respondent who commutes for 30 minutes each day could choose either the second or the third response option, so the options are not mutually exclusive. Because response options overlap, a researcher examining responses to this question cannot differentiate between respondents in adjacent categories. Not providing mutually exclusive response options is a common mistake made when writing survey questions. One could rewrite this survey question to have mutually exclusive response options as "less than 15 minutes; at least 15 minutes but less than 30 minutes; at least 30 minutes but less than 1 hour; 1 hour or more." While a bit wordier, the response options in this revised question are mutually exclusive.

In some cases, as in the previous question, response options are inherently mutually exclusive (only one can be appropriate for any given respondent). In other cases, researchers avoid problems with response options that are not mutually exclusive by asking respondents for the "best" response option or the response option that is highest or lowest on some dimension. For example, asking respondents who did not vote in a recent election, *What is the most important reason why you did not vote in this election: you were too busy, you did not have a strong preference for a candidate, you were ill or did not feel well, or some other reason?*

Providing mutually exclusive response options is one guideline commonly provided for writing survey questions because researchers are typically interested in placing respondents into categories, and violating this guideline makes this categorization impossible. In addition, restricting respondents to select one answer choice when more than one could apply to them is frustrating and confusing for respondents. However, there are some cases in which researchers may want respondents to choose more than one response option, as in a "check-all-that-apply" item. For example, a survey question measuring racial identification may allow respondents to select more than one response option. For example, the earlier question about voter turnout could be rewritten to allow multiple responses: *Why did you not vote in this election? Please select all that are true for you: (1) you were too busy, (2) you did not have a strong preference for a candidate, (3) you were ill or did not feel well, or (4) some other reason?* Responses to these questions can then be transformed for analysis into multiple variables reflecting whether respondents selected each response option.

Allyson Holbrook

See also Check All That Apply; Exhaustive; Response Alternatives

Further Readings

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design*. San Francisco: Wiley.

N

n

The sample size is traditionally labeled *n*, as opposed to the total population size, which is termed *N*. The sample size, *n*, can refer to either the original number of population elements selected into the sample (sometimes called the "designated sample size" or "sampling pool"), or it can refer to the final number of completed surveys or items for which data were collected (sometime called the "final sample size" or "final sample"). In the same vein, it could refer to any number in between such as, for example, the number of elements that have been sampled and contacted but not interviewed. Or it could refer to the number of elements for which complete data are available. Another interpretation or use of the term *n* is the number of elements on the data file and available for analysis.

It is almost always true that *n* is smaller than *N* and usually by orders of magnitude. In fact, the ratio (n/N) is often referred to as the *sampling fraction*. Often the population size *N* is so large relative to *n* that one can safely assume that with replacement sampling holds even if in practice without replacement sampling is implemented. The relative sizes of *n* and *N* also play a role in determining whether the finite population correction factor $[1-(n/N)]$ is sufficiently different from 1 to play a role in the calculation of sampling variance.

Karol Krotki

See also Element; Finite Population Correction (fpc) Factor; *N*; Population; Sample; Sample Size; Sampling Fraction; Sampling Without Replacement

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

N

The total population size is traditionally labeled *N*, as opposed to the sample size, which is termed *n*. The population size *N* refers to the total number of elements in the population, target population, or universe.

N also refers to the number of elements on the sampling frame from which the sample is to be drawn. Since in many cases, the list of population elements contains *foreign elements*, the accurate number of eligible population elements is less than the number of elements on the list. In other cases, the population list not only contains foreign elements but also contains omissions and inaccuracies. These further put into question the validity of the value of *N*, which should be assessed carefully both before and after sample selection and survey implementation.

In some situations *N* is unknown, and in fact one of the objectives of the survey is to estimate *N* and its distributional characteristics. In other situations *N* is known only approximately, and its estimate is refined based on the information obtained from the survey.

It is almost always true that N is larger than n and usually by orders of magnitude. In fact, the ratio (n/N) is often referred to as the *sampling fraction*. Often the population size N is so large relative to n that we can safely assume that with replacement sampling holds even if without replacement sampling is implemented in practice. The relative sizes of n and N also play a role in determining whether the finite population correction factor $[1 - (n/N)]$ is sufficiently different from 1 to play a role in the calculation of sampling variance.

Karol Krotki

See also Element; Finite Population Correction (fpc) Factor; *n*; Population; Sample; Sample Size; Sampling Fraction; Sampling Frame; Sampling Without Replacement

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

National Council on Public Polls (NCPP)

Founded in 1969, the National Council on Public Polls (NCPP) is an association of public opinion polling organizations. Initiated by George Gallup of the American Institute of Public Opinion, the primary goal of NCPP is to foster the understanding, interpretation, and reporting of public opinion polls through the disclosure of detailed, survey-specific information and methods to the general public and the media.

NCPP recognizes that the goal of public opinion polls is to provide reliable, valid, and accurate information. If polls succeed in achieving these goals, scientifically conducted surveys can characterize the public's view on issues, policies, elections, and concerns of the day. But with the enormous amount of polling information available, competing methods of collecting information, and sometimes contradictory results, it is often difficult for the general public and the media to decipher polls that accurately reflect what people think from polls that do not.

NCPP does not pass judgment on specific polls, polling methods, or polling entities but rather advocates that polling organizations whose results reside in the public realm disclose pertinent information about how their surveys are conducted. NCPP maintains that if provided an adequate basis for judging the reliability and validity of poll results, consumers of surveys may assess these studies for themselves.

It is with this goal in mind that NCPP developed a code for member organizations to abide by when reporting survey findings that are intended for or end up in the public domain. These "Principles of Disclosure" include three levels of disclosure, as described on the NCPP Web site.

Level 1 disclosure requires that all reports of survey findings issued for public release by member organizations include the following information, and, in addition, member organizations should endeavor to have print and broadcast media include these items in their news stories:

- Sponsorship of the survey
- Fieldwork provider (if the member organization did not, itself, conduct the interviews)
- Dates of interviewing
- Sampling method employed (e.g., random-digit dialed telephone sample, list-based telephone sample, area probability sample, probability mail sample, other probability sample, opt-in Internet panel, nonprobability convenience sample, use of any oversampling)
- Population that was sampled (e.g., general population; registered voters; likely voters; or any specific population group defined by gender, race, age, occupation, or any other characteristic)
- Size of the sample that serves as the primary basis of the survey report
- Size and description of the subsample, if the survey report relies primarily on less than the total sample
- Margin of sampling error (if a probability sample)
- Survey mode (e.g., telephone/interviewer, telephone/automated, mail, Internet, fax, email)
- Complete wording and ordering of questions mentioned in or upon which the news release is based
- Percentage results of all questions reported

Level 2 disclosure requires member organizations, in response to any specific written request pertaining to any survey findings they have released publicly, to additionally release any of the following:

- Estimated coverage of target population
- Respondent selection procedure (e.g., within household), if any
- Maximum number of attempts to reach respondent
- Exact wording of introduction (any words preceding the first question)

- Complete wording of questions (per Level 1 disclosure) in any foreign languages in which the survey was conducted
- Weighted and unweighted size of any subgroup cited in the report
- Minimum number of completed questions to qualify a completed interview
- Whether interviewers were paid or unpaid (if interview-administered data collection)
- Details of any incentives or compensation provided for respondent participation
- Description of weighting procedures (if any) used to generalize data to the full population
- Sample dispositions adequate to compute contact, cooperation, and response rates

Level 3 disclosure strongly encourages member organizations to do the following:

- Release raw data sets for any publicly released survey results (with telephone numbers and all other identifying personal information removed)
- Post complete wording, ordering, and percentage results of all publicly released survey questions to a publicly available Web site for a minimum of two weeks
- Publicly note their compliance with these Principles of Disclosure

In keeping with its mission, NCPP established the Polling Review Board (PRB) in 1999 as a source for authoritative comment on good and bad practices of public opinion surveys and/or their public dissemination through the media. Comprised of three member organization representatives, the PRB responds publicly to problems or issues of polling practice, presentation, or media coverage. Comments by the PRB on important polling issues are distributed to the media and are available on NCPP's Web site.

PRB members are also available to provide expert insight and answers to polling questions from politicians, the media, or the general public.

Through expert support and educational activities, NCPP works to advance the public's knowledge about how polls are conducted and how to interpret poll results. NCPP has sponsored seminars, workshops, and press conferences in Washington, D.C., and New York City, and publications to promote understanding and reporting of public opinion polls. One such publication is *Twenty Questions a Journalist Should Ask About Poll Results*, by Sheldon Gawiser and Evans Witt, available by request or online on NCPP's Web site. It provides a guide for reporters who cover polls.

NCPP recognizes excellence in reporting of polls through its annual Excellence in Media Coverage of Polls Award. Established in 2002, the award encourages accuracy and insight by professional journalists in communicating poll results to the public. Award recipients have included journalists from *The Los Angeles Times*, the Associated Press, *USA Today*, and ABC News.

The National Council on Public Polls Web site provides an opportunity for poll consumers to interact with polling experts and to follow current debates among polling leaders. It includes information about the council, member organizations, NCPP publications, readings, writings, and presentations by member representatives, and a variety of sources about public opinion surveys.

Lee M. Miringoff and Barbara L. Carvalho

See also Gallup, George; Polling Review Board (PRB)

Further Readings

Gawiser, S. R., & Witt, G. E. (1994). *A journalist's guide to public opinion polls*. Westport, CT: Praeger.

National Council on Public Polls: http://www.ncpp.org

National Election Pool (NEP)

The National Election Pool (NEP) is a consortium of news organizations—ABC, the Associated Press (AP), CBS, CNN, FOX, and NBC—that conducts exit polls, related surveys of voters, and samples of tabulated vote in U.S. elections. These data allow NEP members to project or "call" winners of many political races earlier than would be possible based on final vote count alone. The voter surveys also allow pool members and subscribers to analyze demographic, attitudinal, and other variables that help explain election outcomes.

Typically the exit polls and sample vote count cover top-of-the-ticket statewide races including those for president, U.S. Senate, and governor, as well as selected ballot initiatives. NEP also conducts a national voter survey in general elections. The NEP exit polls are among the largest one-day survey research undertakings anywhere; in the November

2004 elections, approximately 150,000 interviews were conducted in 1,469 U.S. precincts nationwide.

NEP's roots date to 1990. Before then, several television networks fielded their own exit polls and vote count samples individually. In 1990, the broadcast networks ABC, CBS, NBC, and the then-new cable network CNN formed Voter Research & Surveys (VRS) to pool these functions. In 1993, those networks and the Associated Press, a global news network serving newspapers, broadcasters, and more recently online customers, created the Voter News Service (VNS), which merged the VRS exit polling and sample precinct vote count with the National Election Service (NES), a consortium of news organizations that tabulated vote comprehensively on election nights. The cable network, FOX News Channel, joined VNS after the 1996 presidential primaries.

Exit polls are face-to-face surveys of voters as they exit polling places on Election Day. From the time the polls open until about an hour before they close on Election Day, interviewers approach respondents at a systematic interval and ask them to complete self-administered paper questionnaires, which are kept confidential. Samples of voting precincts—stratified by geography and past vote by party—are selected for the exit polls to be representative of the state, or in a national survey, the entire country. In addition to the exit poll sample, a "superset" random sample of precincts is drawn and news stringers (part-time and/or temporary employees) assigned to report vote count as quickly as possible after polls close. As early and absentee voting began to become more widespread in the United States, VNS started supplementing some exit polls with random-digit dial telephone polling the week before the election to reach voters who would not be covered in the Election Day in-person surveys, and these data are incorporated into projections models and analytical survey cross-tabulations.

In the 2000 general election, VNS and its members became enmeshed in controversy over erroneous or premature calls in the presidential race in several states, particularly in Florida—both early on Election Night, based in part on faulty interpretation of the exit polls, and early the next morning, based on faulty interpretation of the vote count models alone. In a congressional hearing in 2001, the VNS partners vowed to improve their systems, and subsequently they hired a contractor to do so, but the computer overhaul failed in the 2002 midterm election and no exit poll or sample precinct vote data were available that night. Thereafter, the VNS members disbanded that organization and formed NEP in its place.

Unlike VNS, the new pool did not have its own staff but hired outside vendors—Edison Media Research and Mitofsky International. Under NEP, Edison-Mitofsky used in essence the same survey and sample precinct methodology as VNS (which Warren Mitofsky and Murray Edelman and others had developed at CBS prior to the formation of VRS) but ran the data through new computer systems. However, NEP abandoned the broader VNS vote count function; the AP, which had maintained its own comprehensive vote count during the VNS era—with stringers collecting vote in statewide and down-ballot races in every county in the country (or towns and cities in the New England states, where official vote is not tallied centrally by counties)—became the sole U.S. source of unofficial vote count. AP vote tabulation data are incorporated into the Edison-Mitofsky projections models when it becomes available on Election Night, helping NEP members call winners in races that were too close to be called from early voter surveys, exit polls, and sample precinct vote count alone.

The first election NEP covered was the California gubernatorial recall in November 2003. NEP covered 23 Democratic presidential primaries and caucuses in early 2004; the general election in all 50 states and the District of Columbia in November of that year; and elections in 32 states in the 2006 midterms.

The pool faced controversy again in the 2004 general election when estimates from exit poll interviews early in the day leaked on Web sites and indicated Democrat John Kerry would win the race for president. Even with more complete samples later in the day, some survey estimates fell outside sampling error tolerances when compared to actual vote. Several hypotheses for the discrepancies were offered, and the pool and Edison-Mitofsky took corrective action, including changes to interviewer recruitment and training procedures and measures to stanch leaks of early, incomplete exit poll data. One of those measures, a quarantine room, was established in 2006 and successfully monitored very closely by NEP, which strictly limited the access that NEP's sponsors could have to the exit poll data on Election Day prior to 5:00 P.M. EST, and this resulted in no early leaks in 2006.

NEP planned to cover 23 states in the 2008 Democratic and Republican presidential primaries and all 50 states plus the District of Columbia in the general election in November 2008. The pool now typically supplements the Election Day exit polls with telephone surveys for early or absentee voters in about a dozen states in a presidential general election.

Michael Mokrzycki

See also Election Night Projections; Exit Polls

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

National Election Studies (NES)

The American National Election Studies (NES) are national surveys of voters in the United States that have been conducted by the University of Michigan before and after every presidential election since 1948. For midterm elections, the NES has conducted post-election studies since 1958. The NES has become the standard bearer for election studies. Indeed, international election studies have patterned their approach and question format after the NES. The popularity of the NES is due, in part, to its consistency. It has asked many of the same questions repeatedly since its inception. This has allowed researchers to develop innovative hypothesis testing through the examination of many variables, which has permitted analysis across people, contexts, and time.

History

The NES grew out of the studies created by the Survey Research Center and the Center for Political Studies of the Institute for Social Research at the University of Michigan. The program always lacked sufficient funding, which limited improvement to the study. The funding that it did receive was primarily used to conduct the survey. As a result, there were rarely changes to the core questions of the study. This also meant that those not directly involved in the program had little influence on the types of questions offered.

In 1977, through the initiative of sociologist Warren E. Miller, the National Science Foundation (NSF) formally established the National Election Studies. With sufficient funding, the NES was expected to fulfill two expectations. First, it was expected to continue the time-series collection of core questions. NSF insisted that they continue collecting data on social background, underlying social and political values, opinions on public policy, political predispositions, participation in the political process, and perceptions of groups, leaders, and political candidates. Second, with NSF funding, the NES was also expected to improve the studies' core concepts and questions.

When the NSF began funding the NES, it mandated that NES become a truly national resource. This meant that researchers at the University of Michigan were expected to seek out and accept suggestions from outside sources, primarily researchers at other institutions. This has granted a greater number of scholars access to the NES, which, in turn, has improved the quality and breadth of the study. The NES research agenda undergoes a great deal of evaluation and revision as the principal investigators, board of overseers, and ad hoc committees all have their say in the direction of each project.

Planning the National Election Studies

Planning for the NES typically begins two years prior to the election to be studied. One year prior to the election, the Survey Research Center at the University of Michigan conducts a pilot study. These pilot studies are designed to test new survey questions, which are typically associated with a special theme or important current events. Usually this means that multiple versions of each question are used and later examined for reliability and validity. All NES questionnaires consist of new questions drawn from the pilot studies and the core time-series questions.

The core time-series questions are selected because they are consistently relevant to national elections, public opinion, and civic participation. These questions are included in the NES to serve two purposes. First, it allows the NES to measure the impact of exogenous shocks to the political system. Second, these time-series allow scholarship to examine the nature and causes of political change more closely.

In addition to their time-series questions, the NES created a specific Senate Election Study to allow researchers to analyze senatorial elections. Since only one third of the Senate's seats are up for election in any election cycle, it has always been difficult for national surveys to sample enough respondents to properly analyze these elections. In 1988, 1992, and 1994, the NES created a special survey that specifically sampled states where Senate elections were taking place. They conducted a similar series of studies associated with the presidential nomination process in 1980, 1984, and 1988. These surveys were designed to understand better how Americans make political choices and learn about politics in multi-candidate arenas that sometimes lack partisan cues.

Conducting the National Election Survey

The NES has traditionally been conducted using face-to-face interviews. There have been instances in which telephone interviewing has been used, but the NES has always returned to face-to-face techniques. In presidential election years, pre-election interviews begin the day after Labor Day and end the day before the election. The post-election interviews begin the day after the election and are usually completed between late December and early January. Midterm election interviews also begin the day after the election and end around the start of the new year. The NES uses a multistage area probability design to create its sample.

Research Opportunity

Unlike many public opinion surveys, the NES has been made available to anyone who wants to use it. A researcher can download the individual responses of each person surveyed since 1948. These data sets are available from Inter-university Consortium for Political and Social Research (ICPSR) or directly from the American National Election Studies Web page. The NES also provides a number of other resources, including technical reports, tables, and graphs. To date, there are more than 5,000 entries on the NES bibliography, demonstrating the wide-ranging research options that are available from analysis of these data.

James W. Stoutenborough

See also Election Polls; Face-to-Face Interviewing; Multistage Sample; Perception Question; Pilot Test; Reliability; Telephone Surveys; Validity

Further Readings

American National Election Studies: http://www.electionstudies.org

Franklin, M. M., & Wlezien, C. (Eds.). (2002). *The future of election studies*. Boston: Pergamon.

Johnston, R. (2000). Capturing campaigns in national election studies. In E. Katz & Y. Warshel (Eds.), *Election studies: What's their use?* (pp. 149–172). Boulder, CO: Westview.

National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Surveys (NHANES) are a group of studies that measure the health and nutritional status of U.S. children and adults. It is conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). NHANES is the only NCHS survey that gathers objective health measurements based on physical examinations. NHANES contributes to the mission of CDC and the Department of Health and Human Services (DHHS) by collecting standardized data that help shape policies and programs to promote health by preventing and controlling disease and disability. Also, NHANES helps NCHS fulfill its responsibility for producing vital and health statistics for the nation.

Background

The NHANES program began in the early 1960s. The first surveys did not have a nutritional component. They were called the National Heath Examination Surveys (NHES). When nutrition assessments were added in the 1970s, the survey name changed to the National Health and Nutrition Examination Survey (NHANES).

The NHES and NHANES surveys were conducted periodically through 1994 and targeted selected age groups. Since 1999, NHANES has been conducted every year and includes people of all ages.

NHANES is a cross-sectional survey with a stratified, multi-stage probability sample design. The NHANES sample is selected from the civilian, noninstitutionalized U.S. population and is nationally representative. NHANES examines about 5,000 persons annually. Participants are selected in 15 counties across the country each year. These data provide an overview of the health and nutrition of the U.S. population at one point in time. NHANES data are also linked to Medicare and National Death Index records to conduct follow-up studies based on mortality and health care utilization.

Data Collection

NHANES consists of three major pieces: (1) health interviews, (2) medical examinations, and (3) laboratory measures. The health interviews take place in the participants' homes. These are conducted face to face, using computer-assisted personal interviewing (CAPI) software on pen-top computers. CAPI was first used in 1992, during the Third National Health and Nutrition and Examination Survey (NHANES III). Before 1992, NHANES interviews were conducted using pencil and paper.

The home interviews are followed by physical examinations. These are done in the NHANES Mobile Examination Centers (MECs). The MEC is made up of four interconnected 18-wheel tractor trailers. Each of the four trailers houses multiple examination rooms. The MEC visit also includes a dietary recall interview and a health interview covering topics too sensitive to ask in the home.

Laboratory specimens, including blood and urine, are also collected in the MEC. Some laboratory tests are conducted on-site, in the MEC laboratory. Others are done at laboratories across the country. Small amounts of urine and blood are also stored for future testing, including genetic testing.

After the MEC examinations, certain subsets of NHANES respondents participate in telephone interviews. All participants receive a report of the results from selected examination and laboratory tests that have clinical relevance.

The topics covered by NHANES vary over time. Because current NHANES data are released in two-year cycles, survey content is modified at two-year intervals. Some topics stay in the survey for multiple two-year periods. When the data needs for a topic are met, it is cycled out of NHANES, and new topics are added. Rotating content in and out over time has several benefits. It gives NHANES the flexibility needed to focus on a variety of health and nutrition measurements. It provides a mechanism for meeting emerging health research needs in a timely manner. This continuous survey design also makes early availability of the data possible.

Release and Use of Data

NHANES data are used to study major nutritional, infectious, environmental, and other chronic health conditions in the United States. The data are used by federal and state government agencies, community health organizations, private industry, consumer groups, and health providers. NHANES is also an excellent resource for secondary data analysis for college students and academic or private researchers.

Since 2000, NCHS has made NHANES public data sets available on its Web site. Most NHANES data are available to the public at no cost. A small number of NHANES data sets are not publicly available because of confidentiality requirements. These few nonpublic data sets can be accessed through the NCHS Research Data Center (RDC). There are some costs associated with using the NCHS RDC.

A growing number of analysts use NHANES data to study major health conditions in the United States. NHANES data users face certain challenges because of the complexity of the survey design and the vast amount of information in NHANES data sets. To address this issue, NCHS and the National Cancer Institute (NCI) developed a Web-based NHANES tutorial. The tutorial was created to meet the needs of NHANES users regardless of their level of experience with NHANES data or their statistical knowledge. The tutorial has also been accredited for earning credits for Continuing Medical Education (CME), Continuing Education in Nursing (CNE), and Continuing Education Units (CEU).

Natalie E. Dupree

See also Complex Sample Surveys; Computer-Assisted Personal Interviewing (CAPI); Cross-Sectional Data; Multi-Stage Sample; National Health Interview Survey (NHIS)

Further Readings

National Center for Health Statistics. (2004, October). *National Health and Nutrition Examination Survey, 2005–2006*. Retrieved February 15, 2007, from http://www.cdc.gov/nchs/data/nhanes/OverviewBrochureEnglish_May05.pdf

National Center for Health Statistics. (2006, August). *Supporting statement for request for clearance, National Health and Nutrition Examination Survey 2007–2008*. OMB No. 0920-0237. Retrieved February 16, 2007, from http://www.reginfo.gov/public/do/PRAViewDocument?ref_nbr=200610-0920-003

National Center for Health Statistics. (2007, February). *Current NHANES Web tutorial*. Retrieved February 15, 2007, from http://www.cdc.gov/nchs/tutorials/nhanes/index.htm

National Health and Nutrition Examination Survey: http://www.cdc.gov/nchs/nhanes.htm

National Health Interview Survey (NHIS)

The National Health Interview Survey (NHIS) is one of a family of health surveys conducted by the National Center for Health Statistics (NCHS), which is the U.S. government's health statistics agency. The NHIS was authorized in 1956 by an act of Congress—the National Health Survey Act—which stipulated that NCHS was "to provide for a continuing survey and special studies to secure accurate and current statistical information on the amount, disruption, and effects of illness and disability in the United States, and the services received for or because of such conditions." NCHS is now part of the Centers for Disease Control and Prevention (CDC), which is part of the U.S. Department of Health and Human Services.

The NHIS is an annual national household survey, conducted throughout the year, of the civilian noninstitutionalized population of the United States. Following a recent sample size reduction due to budget constraints, the annual NHIS sample now consists of approximately 87,500 persons of all ages who reside in approximately 35,000 households. Trained interviewers from the U.S. Bureau of the Census conduct in-person interviews using computer-assisted personal interviewing.

Core Questionnaire and Supplements

Since its inception in 1957, the NHIS has covered a wide range of health topics, including general health status, acute and chronic conditions, use of health care services, health insurance coverage, and disability and its consequences, as well as basic demographic and socioeconomic information. The NHIS questionnaire was substantially revised in 1997, and its stable core now contains three major submodules, which cover (1) the entire family (about whom a knowledgeable adult responds), (2) a randomly sampled child (about whom a knowledgeable adult responds), and (3) a randomly sampled adult (who responds for him- or herself).

The Family Core questionnaire covers everyone in the family, asking about demographics, general health, and health-related topics. It includes a set of age-appropriate questions on activities of daily living (ADLs) and instrumental activities of daily living (IADLs), and questions on cognitive functioning. Health conditions causing these limitations are identified. Other questions deal with use of medical services, medically attended injuries and poisonings, and disability days. Detailed information on health insurance coverage for each family member is obtained.

The Sample Adult Core covers adults ages 18 and over. Topics include functional limitations and selected conditions, such as heart disease, respiratory conditions, diabetes, arthritis and joint problems, and hearing and visual impairments. Other questions cover mental health status and impact, smoking, drinking, and leisure-time physical activity. Questions are asked about usage of health care services, including having a usual place of health care, hospitalizations, and use of doctor and dentist services.

The Sample Child Core roughly parallels the adult questionnaire; in both, the health conditions covered are age appropriate, and in the former, there are additional questions on developmental problems, school-related difficulties, and mental health.

Each year, supplements—additional questions that go into more detail and/or that cover new topics—are sponsored by other government agencies and added to the NHIS. Examples include several supplements on disability, including longitudinal ones, that were fielded in the 1980s and 1990s. Recent supplement subjects have been health promotion, diabetes, cancer, children's mental health, and complementary and alternative medicine. For example, the 2005 Cancer Control Supplement included topics on diet and nutrition, physical activity, tobacco usage, cancer screening, genetic testing, and family history; this was sponsored by the National Cancer Institute, National Institutes of Health (NIH), and the National Center for Chronic Disease Prevention and Health Promotion at CDC. Another example is the 2004 Children's Mental Health Supplement, which contained the Strengths and Difficulties Questionnaire, 32 questions asked of a parent or guardian about the child, sponsored by the National Institute of Mental Health at NIH. NHIS supplements, or variations of them, are often repeated in different years.

Release of Data

NCHS publicly releases NHIS microdata annually from both the core and supplements. Microdata collected during 2004 were released less than 7 months after the end of the data collection year. Currently, all public use files and supporting documentation for data years 1970 through the year of the most recent release are available without charge from the NHIS Web site. Previous years of public use files from 1963 through 1969 will soon be available for downloading from the NCHS Web site as well.

Since data year 2000, NCHS has been releasing quarterly estimates for 15 key health indicators through its Early Release (ER) Program. After each new quarter of data collection, these estimates are updated and then released on the NCHS Web site 6 months after the data collection quarter. The 15 measures covered by ER include (1) lack of health insurance coverage and type of coverage, (2) usual place to go for medical care, (3) obtaining needed medical care, (4) obesity, (5) leisure-time physical activity, (6) vaccinations, (7) smoking and alcohol consumption, and (8) general health status. For each of these health measures, a graph of the trend since 1997 is presented, followed by figures and tables showing age-specific, sex-specific, and race/ethnicity-specific estimates for the new data quarter. Key findings are highlighted. A separate in-depth report on health insurance is also updated and released every 3 months as part of the ER Program. Both quarterly ER reports are released only electronically, on the NCHS Web site.

In addition to releasing NHIS microdata to the public, NCHS staff members publish their own analyses of the data. Series 10 reports provide results of analyses of NHIS data in substantial detail. Among those series reports are three volumes of descriptive statistics and highlights published annually, based, respectively, on data from the NHIS Family Core, Sample Child Core, and Sample Adult Core. NCHS's series Advance Data From Vital and Health Statistics publishes single articles from the various NCHS programs. NCHS's annual report on the health status of the United States (*Health, United States*) contains numerous tables and other analytic results based on NHIS data.

Multiple years of NHIS microdata are periodically linked to other databases, such as the National Death Index and Medicare records. The National Death Index is an NCHS-maintained central computerized index of state death record information. Linkage to the NDI ultimately provides outcome information about underlying and contributing causes of death.

The NHIS also serves as a sampling frame for the Medical Expenditure Panel Survey (MEPS), which was designed to provide policymakers, health care administrators, businesses, and others with information about health care use and costs and to improve the accuracy of their economic projections. It surveys families and individuals, their medical providers, and their employers across the United States. The MEPS families are a subset of those interviewed within the previous year for the NHIS.

When analysis of NHIS data requires access to confidential microdata that are not released publicly, the NCHS Research Data Center allow researchers meeting certain qualifications to access such data under strict supervision. Researchers must submit a proposal for review and approval. Access may be on-site at NCHS or remotely.

Jane F. Gentleman and Susan S. Jack

See also Computer-Assisted Personal Interviewing (CAPI)

Further Readings

Adams, P. F., Dey, A. N., & Vickerie, J. L. (2005). *Summary health statistics for the U.S. population: National Health Interview Survey*. National Center for Health Statistics. Vital Health Stat Series 10, No. 233. Retrieved April 20, 2008, from http://www.cdc.gov/nchs/data/series/sr_10/sr10_233.pdf

Agency for Healthcare Research and Quality, Medical Expenditure Panel Survey: http://www.meps.ahrq.gov

Barnes, P., & Schiller, J. S. (2006). *Early release of selected estimates based on data from the January–June 2006 National Health Interview Survey*. National Center for Health Statistics. Retrieved April 20, 2008, from http://www.cdc.gov/nchs/nhis.htm

Bloom, B., & Dey, A. N. (2006). *Summary health statistics for U.S. children: National Health Interview Survey, 2005*. National Center for Health Statistics. Vital Health Stat Series 10, No. 231. Retrieved April 20, 2008, from http://www.cdc.gov/nchs/data/series/sr_10/sr10_231.pdf

Cohen, R. A., & Martinez, M. E. (2006). *Health insurance coverage: Early release of estimates from the National Health Interview Survey, January–June 2006*. National Center for Health Statistics. Retrieved April 20, 2008, from http://www.cdc.gov/nchs/nhis.htm

National Center for Health Statistics: http://www.cdc.gov/nchs

National Center for Health Statistics *Advance Data* publications: http://www.cdc.gov/nchs/products/pubs/pubd/ad/ad.htm

National Center for Health Statistics annual *Health, United States* publication: http://www.cdc.gov/nchs/hus.htm

National Center for Health Statistics National Death Index: http://www.cdc.gov/nchs/ndi.htm

National Center for Health Statistics Research Data Center: http://www.cdc.gov/nchs/r&d/rdc.htm

National Center for Health Statistics Series 10 reports: http://www.cdc.gov/nchs/products/pubs/pubd/series/sr10/ser10.htm

National Health Interview Survey: http://www.cdc.gov/nchs/nhis.htm

Pleis, J. R., & Lethbridge-Cejku, M. (2007). *Summary health statistics for U.S. Adults, National Health Interview Survey, 2005*. National Center for Health Statistics. Vital Health Stat 10(232). Retrieved April 20, 2008, from http://www.cdc.gov/nchs/data/series/sr_10/sr10_232.pdf

Schoenborn, C. A., Vickerie, J. L., & Powell-Griner, E. (2006, April 11). *Health characteristics of adults 55 years of age and over: United States, 2000–2003*. National Center for Health Statistics. Advance Data 370. Retrieved April 20, 2008, from http://www.cdc.gov/nchs/data/ad/ad370.pdf

National Household Education Surveys (NHES) Program

The National Household Education Surveys Program (NHES) is a series of nationally representative telephone surveys of households in the United States sponsored by the U.S. Department of Education's National Center for Education Statistics. The chief purpose of the surveys is to describe the educational activities and experiences of young children, school-age children, and adults. The NHES program conducts several surveys in three main topic areas: (1) adult education, (2) school-age children's education, and (3) education and care of young children. One of the most widely reported estimates from the NHES is the number of children being homeschooled in the United States. NHES is the only scientific sample survey that regularly produces estimates of the prevalence of homeschooling, estimated in 2003 at 1.1 million U.S. homeschooled children. NHES is also an important source of data about trends in the use of school choice in public schools, revealing that the number of children enrolled in chosen public schools, as opposed to assigned schools, rose from 11 to 15% between 1993 and 2003.

The NHES surveys were first conducted in 1991, and subsequent surveys have been administered in 1995, 1996, 1999, 2001, 2003, 2005, and 2007. Data collections have taken place during the period of January through March or April of these years, and most questions refer to the prior 12 months. All interviews are completed using computer-aided telephone interviewing (CATI). In each survey year, two or more surveys are administered concurrently in order to reduce administration costs. A common screener interview is administered to each sampled household. The NHES screener interview includes a roster of all household members and determines each household member's eligibility to be sampled for one of the extended interviews that is being administered.

Six core NHES surveys have been repeated at least twice and are planned for continuing administration: Early Childhood Program Participation; School Readiness; Parent and Family Involvement in Education; After-School Programs and Activities; Adult Education; and Adult Education for Work-Related Reasons. Other surveys have previously been administered but are not planned to be repeated: Civic Involvement; School Safety and Discipline; and the Household and Library Use Survey. Each year's NHES draws an independent cross-sectional sample; NHES is not a longitudinal study, but time-series analysis is possible because many questions have been repeated in different years.

In each NHES survey, interviews are completed with several thousand individuals. The adult surveys describe the population of civilian, noninstitutionalized adults 16 years of age or older and not enrolled in high school or below. Surveys regarding school-age children and very young children are completed by a knowledgeable adult, usually the child's mother.

Response rates on the NHES surveys have been high relative to most telephone surveys. The response rate on the NHES screener interview in 2005 was 67%. The overall weighted response rate for the Adult Education survey in 2005 was 48%, and the overall weighted response rate for both surveys regarding children in 2005 (Early Childhood Program Participation and After-School Programs and Activities) was 56%. The typical response rate pattern observed in NHES surveys is that surveys asking parents to talk about their children achieve a higher response rate than surveys asking adults to talk about their own education. These rates are achieved by using established techniques to maximize response rates, including sending an advance letter to all sampled households for which a vendor is able to determine a valid mailing address, paying a monetary incentive for participation, making repeated call attempts to each household at different times of day over a period of several weeks, and refusal conversion attempts, where sampled individuals who refuse to participate are asked to reconsider.

Like nearly all sample surveys of the general population conducted by the federal government, the NHES uses complex sampling procedures rather than simple random sampling. This means that the classical approaches to hypothesis testing and the estimation of sampling error and confidence intervals (which assume simple random sampling) are not appropriate for NHES data, as these procedures would generally overstate the precision of the estimates and lead researchers to erroneously conclude that the difference between two estimates is statistically significant when it is not.

Matthew DeBell

See also Advance Letters; Complex Sample Surveys; Computer-Assisted Telephone Interviewing (CATI); Incentives; Refusal Conversion

Further Readings

National Household Education Surveys Program: http://nces.ed.gov/nhes

Nolin, M. J., Montaquila, J., Nicchitta, P., Hagedorn, M., & Chapman, C. (2004). *National Household Education Surveys Program: 2001 methodology report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Princiotta, D., & Bielick, S. (2006). *Homeschooling in the United States: 2003*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Tice, P., Chapman, C., Princiotta, D., & Bielick, S. (2006). *Trends in the use of school choice: 1993–2003*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

National Opinion Research Center (NORC)

The National Opinion Research Center (NORC) is the oldest and largest university-based survey research organization in the United States. It was founded in 1941 at the University of Denver by Harry H. Field. Field was from Britain and had worked for the Gallup Organization and set up Gallup in Britain. Departing from the model of commercial public opinion firms established by Archibald Crossley, George Gallup, Elmo Roper, and others, Field wanted to conduct survey research in the public interest, to serve the nonprofit and government sectors, to improve survey methods, and to advance public opinion research by reviewing and synthesizing results from all organizations. After Field's death in a plane crash in France in 1946, the new director, Clyde Hart, moved NORC in 1947 to the University of Chicago, where it has remained. NORC has played a leadership role in many areas of survey research: organizationally, methodologically, and substantively.

Field organized the first conference ever held in the new field of survey research in Central City, Colorado, in 1946, and this led directly to the founding of the American Association for Public Opinion Research (AAPOR) and the World Association for Public Opinion Research (WAPOR) in 1947.

NORC researchers have pioneered in studying the error structures of surveys and developing methodologies to improve survey quality. These efforts include Herbert Hyman's work in the 1950s on interviewer effects, Norman Bradburn's studies on the measurement of psychological well-being, Bradburn and Seymour Sudman's research on response effects, the studies of context effects by Kenneth Rasinski, Tom W. Smith, and Roger Tourangeau, and the studies conducted of employers, congregations, and voluntary associations using hypernetwork sampling from the General Social Survey (GSS).

NORC has also conducted seminal research in many areas. In 1942, it conducted the first national survey on race relations; this led to a long series on intergroup relations. In 1947, the first national study of occupational prestige was carried out. Measures of occupational prestige were then refined and updated in 1963–65 and in 1989 as part of NORC's GSS. In 1963, immediately following the death of President John F. Kennedy, the Kennedy Assassination Study was fielded. In 2001, in the aftermath of the September 11 terrorist attacks, NORC conducted the National Tragedy Study, drawing on many questions from the Kennedy Assassination Study and from the GSS. In 1970, for the Kinsey Institute, NORC carried out the first national survey to measure many aspects of sexual behavior, including homosexuality. On the 1985

GSS, the first national, egocentric, social network study was completed. In 1985–86, in Chicago, NORC conducted the first systematic probability sample of the homeless. In 1998, the first national sample of congregations was fielded.

Presently NORC has multiple offices in both the Chicago and Washington, D.C., areas. It is divided into three divisions: (1) administrative, (2) survey operations, and (3) academic centers. The administrative division covers basic management functions such as accounting and human resources.

The survey operations division designs and conducts data collection efforts. It is divided into several research departments along substantive lines: Economics, Labor, and Population; Education and Child Development; Health Survey, Program, and Policy Research; Information Technology; International Projects; Statistics and Methodology; and Substance Abuse, Mental Health, and Criminal Justice. Most frequently these departments carry out complex, large-scale, in-person surveys under contract with the federal government.

The academic division consists of several research centers: the Alfred P. Sloan Center on Parents, Children, and Work; the Center on the Demography and Economics of Aging; the Data Research and Development Center; the Ogburn-Stouffer Center for the Study of Social Organization; and the Population Research Center. These centers work with the research departments in designing surveys, conduct some surveys themselves, and analyze results from NORC surveys and other data sources.

One area of special concentration at NORC is panel studies. Over the years these have included such projects as the Midtown Manhattan Study, High School and Beyond, the old and new cohorts of the National Longitudinal Survey of Youth, the National Educational Longitudinal Study, and Baccalaureate and Beyond.

A second area of specialization is studies of societal change. In early years, these included surveys for the Department of State on foreign policy issues and trends on anti-Semitism and race relations. Since 1972, the GSS has monitored societal change with 26 nationally representative surveys and more than 1,000 time series.

A third area has been cross-national and comparative studies, including the Civic Culture Study in 1959, the Soviet Interview Project in 1980, the GSS-related International Social Survey Program from 1985 to the present, and the recent Qatar education project.

A final example of an area of concentration involves the establishment of professional standards for the field of survey research. As noted above, NORC was instrumental in establishing AAPOR and WAPOR. More recently NORC personnel played central roles in the adoption by AAPOR and WAPOR of *Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys*, the work of several National Academies of Science panels, and the formulation of the rules of the International Organization for Standardization for market, opinion, and social research.

NORC's work is very varied and covers many other areas are well. Other examples include the Florida Ballots Project, which counted and analyzed all contested ballots in the 2000 Florida general election; annual rankings of America's best hospitals, which identified the nation's top hospitals by specialty; the National Social Life, Health, and Aging Project, which examined the sexual behavior of older Americans; and Poetry in America, which studied exposure to this literary form.

Tom W. Smith

See also American Association for Public Opinion Research (AAPOR); General Social Survey (GSS); World Association for Public Opinion Research (WAPOR)

Further Readings

Bova, P., & Worley, M. P. (1991). *Bibliography of publications, 1941–1991: A fifty year cumulation*. Chicago: National Opinion Research Center.

Converse, J. M. (1987). *Survey research in the United States: Roots and emergence, 1900–1960*. Berkeley: University of California Press.

Hackett, J. (1992). *America by number: NORC Report 1991*. Chicago: National Opinion Research Center.

National Opinion Research Center: http://www.norc.org

Smith, T. W. (2007). The Midwest and the rise of survey research. In R. Sisson, C. Zacher, & A. Cayton (Eds.), *The American Midwest: An interpretive encyclopedia* (pp. 1723–1724). Bloomington: Indiana University Press.

Network Sampling

Network sampling is widely used when rare populations are of interest in survey research. Typically, sampling frames do not exist for rare populations because usually there is little information on the size and magnitude of the population. Two main methods can be employed in a survey with an unknown population: *screening* and *salting*. The first way is to screen for respondents of interest during the interview, and the second approach is to acquire sample units through official records or documents. Both approaches have shortcomings. Screening requires high costs. Salting entails difficulties with obtaining membership information, because official personnel records may be confidential. Network sampling is considered an alternative to the previous ways of estimating rare populations in which sampling frame is almost impossible to obtain.

Network sampling is also called *snowball sampling* or *multiplicity sampling*. This sampling technique is widely used to estimate populations such as the homeless, Korean War veterans, and patients with rare forms of cancer. Particularly, it has been found that network sampling was much more efficient than other conventional sampling methods for estimating the number of cancer patients. Most network samples have employed family members, relatives, and friends as informants; network informants report about all persons in their network; and sizes vary to degree from respondent to respondent.

Indeed, obtaining an initial sampling frame is a prerequisite for any network sampling method, and the quality of that frame is essential for the ultimate success of the method. Although family members or relatives are used as the network for many of the sampling frames, it need not be restricted to them, depending on the topic of a study. For instance, let us suppose that Korean Americans living in Michigan are the target population of a study. Network sampling using families, relatives, friends, and even casual associates may be useful for this case. In addition to blood kinship, membership lists can be used: Korean religion membership lists, Korean association lists, and so on. That is to say, using multiple sources included in the network sampling increases the network frame in quality and coverage.

To estimate the unbiased survey statistics, weighting is necessary for network sampling; the total eligible respondents of a particular network are weighted by the reciprocal of one over the total number of the particular network. Generally, interviewer costs are a primary concern for the network sampling. In this method, interviewers should meet with potential respondents who were identified by informants to see whether the respondents are eligible for a particular survey. This process increases interviewer time and costs, though both depend largely on the size of network. However, this sampling reduces screening costs.

Geon Lee

See also Multiplicity Sampling; Respondent-Driven Sampling (RDS); Snowball Sampling

Further Readings

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Sudman, S., & Freeman, H. E. (1988). The use of network sampling for locating the seriously ill. *Medical Care*, *26*(10), 992–999.

New York Times/CBS News Poll

The *New York Times*/CBS News poll was the first newspaper–television polling partnership between two major national news organizations in the United States and was launched with a nationwide telephone survey of nearly 800 adults in early November 1975.

On November 4, 1975, the *CBS Evening News with Walter Cronkite* aired a report regarding the American public's attitudes toward President Ford and his response to the possible default of New York City. The following morning, "Poll Finds Public Thinks Ford Minimizes City Peril," an article by Robert Reinhold, was on page 1 of *The New York Times*.

That first poll evolved from discussions between Henry R. Lieberman, Assistant to the Executive Editor of *The New York Times*, and Warren Mitofsky of CBS News. It was followed by an agreement between the two organizations to conduct a series of monthly national telephone surveys and primary election day exit polls to cover the 1976 presidential election campaign. Both the *Times* and CBS News wanted access to their own political polling in order to add greater dimension to their 1976 political coverage and an independent news stream of exclusive polling data.

The partnership has endured and flourished over the years for a number of reasons. Newspapers and television network news are not in direct competition with each other. Each organization's expertise and their different needs enhance the end result. The partnership saves both organizations money. By sharing the work and the results, a poll essentially costs each partner half as much as a solo effort and guarantees two sets of eyes on every aspect of the polling operation.

That first contract worked out the long-standing agreements about the poll's name and when it would be released. In the paper, the poll is identified as the "*New York Times*/CBS News poll." On the CBS News broadcasts, it is the "CBS News/*New York Times* poll."

A joint poll is usually released first on the *CBS Evening News*, broadcast in the East at 6:30 p.m. At that time, CBS also releases the poll to their network radio and Web site. Their press release for the poll is then emailed to a wide audience, including the news wires and other media outlets.

The *Times* publishes the poll article in the paper the following morning. But, after 6:30 p.m. the evening before, the *Times* is free to post the poll story on its Web site. Some poll stories are also printed in *The International Herald Tribune*. The full question wording and results are also posted on both newspapers' Web sites.

Initially, responsibilities were divided for practical reasons. CBS already had an election unit in place, with statisticians and computer programmers, and so provided the sampling, weighting, and technical expertise.

From the beginning, the *Times* and CBS News handled their own fieldwork and continue to do so. The *Times* is in charge of hiring and training the interviewers and managing the data collection. When the surveys were conducted on paper, the interviewing was conducted on West 43rd Street in the advertising department of the *Times*—a large room with desks, telephones, and an advertising staff that cleared out by 5:00 p.m. and did not work weekends. Desks for weekday interviewing were located throughout the newsroom.

The introduction of CATI interviewing in 1991 necessitated the relocation of the interviewing operation to the CBS offices on West 57th Street. Currently, there is a dedicated survey room in the CBS Broadcast Center, with space for about 50 interviewers with monitoring capability and supervisor stations. The *Times* remains responsible for hiring and training the interviewers and maintaining records. But, as with many aspects of the *Times/*CBS News poll, the organizations work together on staffing issues.

There have been changes in the methodology and procedures over time. About the same time as the interviewing was switched to CATI technology, the sampling changed from Mitofsky-Waksberg sampling to list-assisted sampling. The weighting program has been adjusted over time. Some exit polls conducted by CBS News were joint projects with the *Times* until the major television networks first joined forces to conduct exit polls in 1990.

Both the *Times* and CBS News have small departments that create the questionnaires, manage the data collection, and analyze the poll results. The CBS News Election and Survey Unit works directly with executive producers and producers of the *Evening News*, *60 Minutes, The Early Show*, radio, and the Web. The News Surveys Department of the *Times* works directly with the national editor, the Washington Bureau, the Foreign and Metro News desks, and other department heads.

Teams from the *Times* and CBS News develop the questionnaire together, with each bringing subjects and questions to the table, often after consulting with reporters, editors, and producers. Usually the *Times*/CBS polls deal with national politics and policy, but polls often also contain questions on other topics, including business, sports, travel, and culture. Occasionally, polls are conducted with samples of respondents other than national adults, including state and local surveys and polls of convention delegates, business executives, and teenagers.

Although the questionnaire design and data collection are joint operations, the *Times* and CBS News go their separate ways once the survey is completed. Each organization receives tables with banners of standard variables and has access to an interactive system for generating custom tables. Every poll receives two simultaneous but independent analyses by separate teams. That can, and sometimes does, lead to different emphasis in the resulting broadcasts and articles.

Through the decades of collaboration, each side may (and does) conduct polls outside the partnership, often because of lack of interest in a specific polling topic or issue or an outlet by one of the partners. For example, polls in the New York metropolitan area or polls for special series in the *Times* are frequently conducted by the *Times* without CBS. CBS often does surveys without the *Times* for its own special broadcasts.

After more than 30 years interviewing about a half-million respondents in nearly 450 surveys, the partnership is still going strong.

Marjorie Connelly

See also List-Assisted Sampling; Media Polls; Mitofsky-Waksberg Sampling; Random-Digit Dialing (RDD)

Further Readings

Kagay, M. R. (1991). The use of public opinion polls by *The New York Times:* Some examples for the 1988 presidential election. In P. J. Lavrakas & J. K. Holley (Eds.), *Polling and presidential election coverage*. Newbury Park, CA: Sage.

Kagay, M. R. (1995). The evolving use of public opinion polls by *The New York Times:* The experience in the 1992 presidential election. In P. J. Lavrakas, M. W. Traugott, & P. V. Miller (Eds.), *Presidential election polls and the news media*. Boulder, CO: Westview.

Kagay, M. R. (2000). Continuing evolution in the use of public opinion polls by the *New York Times:* The 1996 presidential election experience. In P. J. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy*. New York: Chatham House/CQ Press.

Neyman Allocation

Stratified samples are commonly used when supplementary information is available to help with sample design. The precision of a stratified design is influenced by how the sample elements are allocated to strata. Neyman allocation is a method used to allocate sample to strata based on the strata variances and similar sampling costs in the strata. A Neyman allocation scheme provides the most precision for estimating a population mean given a fixed total sample size.

For stratified random sampling, the population is divided into H mutually exclusive strata. In each stratum, a simple random sample is drawn without replacement. Neyman allocation assigns sample units within each stratum proportional to the product of the population stratum size (N_h) and the within-stratum standard deviation (S_h), so that minimum variance for a population mean estimator can be achieved. The equation for Neyman allocation is

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} n,$$

where n_h is the sample size for stratum h and n is the fixed total sample size. The effect of Neyman allocation is to sample more heavily from a stratum when (a) the population size of the stratum is large; (b) the variability within the stratum is large, so that the heterogeneity needs to be compensated.

Of note, Neyman allocation is a special case of optimal allocation whose objective in sample allocation is to minimize variance of an estimator for a population mean for a given total cost. It is employed when the costs of obtaining sampling units are assumed to be approximately equal across all the strata. If the variances are uniform across all the strata as well, Neyman allocation reduces to proportional allocation where the number of sampled units in each stratum is proportional to the population size of the stratum. When the variances within a stratum are different and are specified correctly, Neyman allocation will give an estimator with smaller variance than proportional allocation.

The major barrier to the application of Neyman allocation is lack of knowledge of the population variances of the study variable within each stratum. In some situations, historical estimates of strata variances can be used to provide good approximation to Neyman allocation for the current survey sample. For example, the Medical Expenditure Panel Survey Insurance Component (MEPS IC) is an annual survey of establishments that collects information about employer-sponsored health insurance offerings. To implement Neyman allocation, stratum variance estimates were obtained from the 1993 National Employer Health Insurance Survey for the initial MEPS IC 1996 and later from prior MEPS IC surveys.

In situations where estimated population variances within each stratum are not easily available, an alternative is to find a surrogate variable (a proxy) that is closely related to the variable of interest and use its variances to conduct a Neyman allocation. For example, the U.S. Government Accountability Office conducted a survey in 2004–2005 to estimate the average and median purchase prices of specified covered outpatient drugs (SCODs) in a population of 3,450 hospitals. Since a direct measure of purchase prices for SCODs was not available at the time of sample selection, the total hospital outpatient SCOD charges to Medicare was used as a proxy to carry out the Neyman allocation.

In practice, Neyman allocation can also be applied to some selected strata instead of all strata, depending on specific survey needs. For example, the National Drug Threat Survey 2004 was administered to a probability-based sample of state and local law enforcement agencies. The sample frame of 7,930 law enforcement agencies was stratified into a total of 53 strata. Of those 53 strata, 50 strata were formed based on the geographic locations of the local law enforcement agencies. A Neyman allocation was used to allocate sample to these strata. The remaining three strata were constructed to represent specific groups of state and local law enforcement agencies, including all state-level and large local law enforcement agencies. To ensure a thorough analysis of the domestic drug situation, these three strata were sampled with certainty.

Ranked set sampling (RSS) is another sampling protocol that can benefit substantially from the implementation of Neyman allocation. In RSS, the various rank order statistics serve the role of strata in a stratified sampling approach. Neyman allocation in RSS assigns sample units for each rank order statistic proportionally to its standard deviation. That is,

$$n_h = \frac{S_h}{\sum_{h=1}^{H} S_h} n.$$

Here, H refers to the total number of rank order statistics and S_h denotes the standard deviation for the h^{th} rank order statistic.

Haiying Chen

See also Optimal Allocation; Proportional Allocation to Strata; Ranked-Set Sampling; Stratified Sampling

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, *20*, 1–31.

900 Poll

A 900 poll is a one-question unscientific "survey" that typically is taken by having television viewers or radio listeners call into a 1-900-number that *involves a cost to the caller*—sometimes a considerable cost. A different 900-number is given for each "response" that the poll allows the self-selected respondents to choose as their answer to whatever the survey question is. These polls are typically sponsored over a brief period of time—often an hour or less, for example, within a television program or shortly after it ends. For example, callers who prefer Contestant A (or Position A on an issue) and those who prefer Contestant B (or Position B on an issue) use separate 900-numbers. It is possible to offer callers more than two answer choices, and thus more than two 900-numbers, but typically these polls utilize only two or three choices.

Such polls have no scientific standing because there is no way to know what target population is represented by those who choose to dial in. Since this is a nonprobability sample, there is no valid way to calculate the size of the sampling error. Additional threats to their validity include the possibility that the same person will call in more than once.

Nonetheless these polls offer a vehicle for media organizations to provide their audience with a feeling of involvement in the programming, since the poll results are typically reported during the show and/or used to make some decision as part of the programming—for example, who won the competition. They also can serve as a source of revenue for the organization that conducts them, and, depending on how much is the charge to call in and how many people respond, they can generate a good deal of profit as they are relatively inexpensive to run.

Paul J. Lavrakas

See also 800 Poll; Nonprobability Sampling; Self-Selected Listener Opinion Poll (SLOP)

Nominal Measure

A *nominal measure* is part of taxonomy of measurement types for variables developed by psychologist Stanley Smith Stevens in 1946. Other types of measurement include ordinal, interval, and ratio. A nominal variable, sometimes referred to as a *categorical variable*, is characterized by an exhaustive and mutually exclusive set of categories. Each case in the population to be categorized using the nominal measure must fall into one and only one of the categories. Examples of the more commonly used nominal measures in survey research include gender, race, religious affiliation, and political party.

Unlike other types of measurement, the categories of a variable that is a nominal measure refer to discrete characteristics. No order of magnitude is implied when comparing one category to another. After the relevant attributes of all cases in the population being

Table 1 Example of three types of descriptive statistics appropriate for nominal measures

	Count	*Proportion*	*Percentage (%)*	*Ratio (Males to Females)*
Male	651	0.484	48.4	0.938
Female	694	0.516	51.6	
TOTAL	1345	1.000	100.0	

measured are examined, the cases that share the same criteria are placed into the same category and given the same label, for example, "Female" or "Male."

Numbers can be used as labels, but great care should be used when using the variable in statistical analyses. The number assignment in place of a more descriptive label is completely arbitrary. Because the categories of a nominal variable are without mathematically measurable relationship to each other, there is no measure of standard deviation to apply to such a measure. As a result, the types of statistical analysis that can be used with such variables are limited. The only appropriate measure of central tendency is the mode; the mean or median of such a variable is meaningless.

For each of the categories of a nominal variable, one can calculate a proportion, a percentage, and a ratio. The proportion would be the number of cases having the selected value of the variable divided by the total number of cases resulting in a value of zero (none of the cases), one (all of the cases), or a value in between. The percentage for the same category would simply be the proportion multiplied by 100. The ratio is a measure of two categories of the variable in relation to one another. Ratios are calculated by dividing one category by another category. Table 1 illustrates these three types of descriptive statistics appropriate for nominal measures.

Measures of the strength of the relationship between two nominal variables, often called *contingency tests*, can be calculated using a chi-square test, which compares the observed counts in each category to the expected values if there were no relationship. The Fisher's Exact test is appropriate when both nominal variables are dichotomous (have only two values). A variety of other nonparametric tests are available that are appropriate for a variety of situations, including empty cells in a cross-tabulation of two nominal variables, sensitivity to extremely large marginal counts, and other factors that can disturb the underlying assumptions of the more commonly used chi-square and Fisher's Exact tests.

James Wolf

See also Chi-Square; Contingency Table; Interval Measure; Level of Measurement; Ordinal Measure; Ratio Measure

Further Readings

Stevens, S. S. (1946, June 7). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Nonattitude

Nonattitude refers to the mental state of having no attitude or opinion toward some object, concept, or other type of stimulus. In survey research, this is manifested by an overt *no opinion* or *don't know* response to an attitude question, but it may also be hidden by a random or guesswork choice of answers to avoid appearing ignorant. Additionally, it is likely that not all *no opinion* or *don't know* responses reflect nonattitudes. This makes it hard to estimate how many respondents have nonattitudes toward the object.

How the Problem Was Uncovered

The nonattitude problem became prominent when the National Election Survey (NES) reinterviewed panels of Americans at 2-year intervals in the 1950s. Political scientist Philip Converse observed the low *stability* of individuals' answers given 2 years apart on issues that had been widely discussed by political leaders and the media. Question reliabilities ranged from .23 to .46. He also noted a lack of constraint or structure in responses to different policies: Most people did not consistently choose liberal or conservative policies within a single survey. The mean correlation between domestic policy responses was .23. A survey of congressional candidates of the two parties, interviewed with the same questions, found a mean interitem correlation of .53. Later NES panel studies of political elites showed that their responses had much more reliability over time as well as much greater interitem correlation. These findings were confirmed by panel studies in the 1970s and surveys since in many countries.

Converse concluded that a great many people had no opinions on major issues of the day and were concealing this by randomly choosing responses rather than answer "Don't know," "Undecided," or "No opinion" even when these alternatives were offered in a nonjudgmental manner. The observed (low) correlations over time and between issues could be produced by one stratum holding real opinions, which were highly stable and consistent, and another stratum of covert nonopinion-holders expressing pseudo-opinions. Assuming no real attitude change over the 2-year periods, he estimated the percentage of covert nonopinion-holders on each question from the number of changed answers, added in the overt nonopinions, and argued that from 20% to 80% of the public had nonattitudes on a wide range of policy questions. This cast doubt on the meaning of most reported opinion survey results, and on the ability of much of the public to form meaningful opinions on the political issues of the day and influence elite decision making. It also led to a major methodological, theoretical, and ideological controversy.

Alternative Models With Latent Attitudes

Critics analyzing the same data rejected the idea that a large part of the public had nonattitudes on leading public issues. Alternative theories to explain the observed instability and incoherence of responses include the following:

1. Measurement error produced by vague and ambiguous questions, concealing real attitudes, which could be revealed by better questions
2. The influence of temporary stimuli—events in the news or in personal life—leading to wide variations in momentary feelings around underlying attitudes
3. The possibility that each object has a variety of elements or considerations about which the individual has positive or negative feelings, but "samples" unsystematically in answering the questions—perhaps randomly, perhaps in response to recent events or cues given by question wording or sequence
4. Those who more systematically inventory the considerations they hold in mind may have a near balance of positive and negative feelings—an ambivalence making their answers unstable from time to time or under different question wordings, although they have strong feelings about the issue.

Critics of the nonattitude hypothesis have used structural equation models to show that the pattern of observed (low) correlations could be the result of most people having stable underlying attitudes, albeit very weakly connected to their responses to the particular questions. According to some estimates, these *latent attitudes* were quite stable, with correlations over 2-year periods ranging from .8 to .9. Instead of a public made up of people with attitudes and people without, public responses in a particular issue area might come from a *latent continuum* of attitude holding, ranging from those with highly reliable and interrelated opinions (such as those found in elites), through those with general pre-dispositions producing only modest degrees of reliability and structure, to a residue with total absence of attitudes, admitted or concealed. Another model uses the idea of *issue publics*—that there are small groups of people with stable, organized ideas in particular issue areas but with only loose underlying attitudes, or none at all, toward policies in other areas. The rest of the public may have poorly structured attitudes in all of the areas, or nonattitudes in some or all. Because political elites have to inform themselves, discuss, and take stands on a wide range of issues, they develop consistent attitudes, based on a general ideology or party loyalty linking many issues.

All these alternative models find stable underlying attitudes in the public at the expense of admitting that public responses to specific policy questions are unstable and only loosely connected to real attitudes. Since the same questions produced stable and coherent answers from political elites and were worded in the terms found in the political discourse of the media, candidates, and decision makers, the large error terms calculated for the questions can be interpreted as reflecting the weakness of public opinion, not the survey technique.

Do Nonattitudes Matter?

A public poorly equipped to relate its underlying attitudes to current policy issues would seem little more likely to have a strong influence on policy than one with nonattitudes. Benjamin Page and Robert Shapiro counter that the nonattitudes or weakly connected attitudes to specific issues do not cripple democracy, because *collective public opinion*, the aggregate of favorable and unfavorable attitudes toward policies and candidates, is rather stable and changes rationally to respond to social and economic problems. They admit imperfections in the process, including failures of the information-providing system, elite misleading or manipulation, and the biasing effects of economic inequality on the "marketplace of ideas," but consider the past 50 years of American experience as evidence that public opinion matters. Comparative research across societies is needed to show which kinds of parties, media institutions, and social organization do better at overcoming the nonattitude problem and improve the correspondence of policies with the public's interests and values.

How Surveys Can Deal With the Nonattitude Problem

Remedies for the survey researchers' nonattitude problem are of several kinds:

1. Screening questions can cut down the number of pseudo-opinions that obscure the picture of actual public opinion.
2. Multi-item scales within issue areas reduce reliance on unreliable single questions. They allow factor and latent attitude analysis to identify underlying attitude dimensions and test the extent to which these dimensions are related to particular policy or candidate choices.
3. Given enough items, the consistency of respondents' attitudes can be measured by the spread of item responses around the respondent's mean position. Using intercorrelation of items to measure attitude constraint at the group level can be misleading if the group has low variance; low intercorrelations may result from high consensus rather than nonattitudes.
4. Nonopinions that slip through the screening questions can be detected by asking questions that reveal contradictory answers and open-ended probes that reveal empty responses.
5. Ideally, the same people should be reinterviewed, preferably several times over a period of years, to check on stability of answers and underlying attitudes and to distinguish stable attitude change from measurement error and weak attitudes.
6. For some purposes, researchers may be interested in what people's attitudes would be, or whether nonattitudes would be replaced by attitudes, if they were exposed to new information or arguments. One can postulate a universe of *potential opinion response* under different conditions and set up survey experiments to sample from that universe. Every opinion poll is an "experiment" on how people respond to certain formulations of issues, given their exposure to certain experiences, information, recent news, and guidance by opinion leaders. What people were actually thinking before the interviewer arrived can only be approximated by surveys. What they would think if better informed or encouraged to deliberate more seriously may also be worth trying to approximate.

Allen H. Barton

See also Attitude Measurement; Attitudes; Cognitive Aspects of Survey Methodology (CASM); Deliberative Poll; Don't Knows (DKs); Measurement Error; Reliability; Validity

Further Readings

Converse, P. E. (1970). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.

Page, B. I., & Shapiro, R. Y. (1992). *The rational public*. Princeton, NJ: Princeton University Press.

Saris, W. E., & Sniderman, P. M. (Eds.). (2004). *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change*. Princeton, NJ: Princeton University Press.

Smith, T. W. (1984). Nonattitudes: A review and evaluation. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 215–255). New York: Russell Sage Foundation.

Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge, UK: Cambridge University Press.

Noncausal Covariation

Although correlation is a necessary condition for causation, it is not a sufficient condition. That is, if X and Y can be shown to correlate, it is possible that X may cause Y or vice versa. However, just because correlation is established between the two variables, it is not certain that X causes Y or that Y causes X. In instances when X and Y are correlated but there is no empirical evidence that one causes the other, a researcher is left with a finding of *noncausal covariation*. A researcher can speculate that one variable causes the other, but unless there is empirical evidence demonstrating an internally valid casual relationship, the researcher has no solid ground upon which to claim the relationship is causal.

In survey research, researchers rarely have valid evidence upon which to base conclusions of causation. Many researchers forget this and often interpret and report their results as though a causal relationship does exist between variables. For example, a researcher may find a correlation between minority status and the willingness to cooperate in a survey when sampled. However, merely finding that minority status is correlated with someone's response propensity is not sufficient to claim that being a racial or ethnic minority person "causes" one to be less likely to participate in surveys. Instead, it is likely that some other variables that are correlated with both being a minority and not being as willing to participate in surveys, such as educational attainment, are the real causal agents.

To demonstrate a causal relationship using a research design with strong internal validity, a true experiment is necessary. Experiments require that random assignment of respondents be carried out with exposure to different levels of the independent variable that the researcher controls. Then, in its simplest form, the experiment will show whether the group assigned to one level of the independent variable shows statistically different levels of the dependent variable than does the group exposed to the other level of the independent variable. If it does, then a causal relationship has been identified. For example, if survey respondents were randomly assigned to one of two levels of prepaid incentives ($5 or $10), then the researcher could determine whether the difference in incentives changed the response rate of the group getting the higher incentive. If it did, then the researcher has evidence of causation, not merely correlation.

Unfortunately, there are many relationships that survey researchers are interested in studying that do not readily lend themselves to experimentation. Although there are other statistical techniques that can be used to investigate whether a correlational relationship is likely to also represent a casual relationship, without an experimental design, a researcher cannot be as confident about drawing cause-and-effect conclusions and often must resign herself or himself to acknowledging that the relationship is one of noncausal correlation.

Paul J. Lavrakas

See also Dependent Variable; Experimental Design; Independent Variable; Internal Validity; Random Assignment

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton-Mifflin.

Noncontact Rate

The noncontact rate for a survey measures the proportion of all sampled cases that are *never contacted* despite the various efforts that the researchers may set in motion to make contact. By default, if a sampled case is never contacted, then no original data for the survey can be gathered from it, other than observations an in-person interviewer might make of the housing structure or neighborhood. For surveys in which the initial sampling unit is a household or business and then there is a respondent sampled within that unit, a noncontact rate can be calculated both at the unit level and at the within-unit (respondent) level. In theory, a noncontact rate of zero (0.0) means that every eligible sampled case was contacted, whereas a noncontact rate of one (1.0) means none of the sampled eligible cases were contacted. Neither of these extreme conditions is likely to occur in a survey. However, the best of commercial, academic, and government surveys in the United States achieve noncontact rates of less than 2%, meaning that more than 49 of every 50 eligible sampled cases are contacted at some point during the field period.

In face-to-face and telephone surveys of the general public, businesses, or specifically named persons, noncontacts result from no human at a household or business ever being reached by an interviewer during the survey's field period, despite what is likely to be many contact attempts across different days of the week and times of the day or evening. In mail and Internet surveys, noncontacts result from the survey request never reaching the sampled person, household, or business due to a bad address, transmittal (delivery) problems, or the person never being at the location to which the survey request is sent during the field period.

Calculating a noncontact rate is not as straightforward as it may first appear due to the many sampled cases in almost all surveys for which the researcher is uncertain (a) whether they are eligible and/or (b) whether they really were "contacted" but did not behave in such a way that provided the researcher with any certainty that contact actually occurred.

Unit-Level Noncontact Rate

At the unit level (household or business), the numerator of the noncontact rate can be computed by tallying up all those sampled units for which the researchers are certain contact attempts were made. In addition to these cases, the researchers must make an informed decision about what portion of the units for which it is uncertain if contact was made also should be included in the numerator (and the denominator). This uncertainty differs when the survey is interviewer-administered versus when it is done via mail or Internet. In the case of in-person surveys, interviewers who approach homes or businesses can make informed judgments about whether the unit looks to be occupied. If it is determined to be occupied and no contact is ever made with an occupant, then that unit must be included in the numerator (and the denominator) of the noncontact rate calculation. If it is determined the unit is not occupied then that unit is not counted as a noncontact and thus not counted in the numerator (and may or may not be counted in the denominator depending on whether all sampled cases or only "eligible" cases are included in the denominator). This estimation of eligibility (referred to as *e*) of which additional units to count as noncontacts is further complicated when a survey has unusual eligibility criteria (e.g., only adults aged 35–49 years), because some of the apparent noncontacts would have been found to actually be ineligible had contact been made. In these surveys, the researchers must make informed (and defensible) decisions about how to estimate which of these cases should be included in the numerator of the unit-level noncontact rate calculation. The denominator of the unit-level noncontact rate can be calculated either by including all cases, or by including all known eligible cases, or by including all known eligible cases plus an estimate (*e*) of the portion of unknown eligibility cases that are judged to be eligible. As noted above, for mail and Internet surveys, sampled cases from which there is no reply whatsoever to the researchers are very difficult to classify as to whether contact ever was made. Again, the researchers need to make a reasonable judgment about what proportion of these cases should be counted as eligible and what portion of these should be counted as being implicit refusals rather than as noncontacts in the noncontact rate calculation. Any of these cases that are counted as refusals should *not* enter into the noncontact rate numerator.

Respondent-Level Noncontact Rate

When a survey samples a specific respondent within a unit, then a respondent-level noncontact rate also can be calculated. The respondent-level rate differs from the unit level in that some of the contacted units will end the field period as a noncontact at the respondent level, but not at the unit level. That is, even though contact has been made with some other person at the home or business, no contact is ever made with the selected respondent. Because of this, the respondent level noncontact rate will almost always be higher than the unit-level rate and will never be lower. Apart from this, the considerations that apply when calculating a unit-level noncontact rate are essentially the same ones that apply when calculating the rate at the respondent level.

Paul J. Lavrakas

See also *e*; Eligibility; Field Period; Noncontact; Standard Definitions; Unit Level

Further Readings

AAPOR. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Retrieved April 21, 2008, from http://www.aapor.org/uploads/standarddefs_4.pdf

Noncontacts

Noncontacts are a disposition that is used in telephone, in-person, mail, and Internet surveys both as a temporary and a final disposition. Two primary types of noncontacts can occur in surveys. The first type occurs when a researcher makes contact with a household or other sampling unit, and no one is present to receive the contact. The second type of noncontact occurs when a researcher makes contact with a household or other sampling unit, but the selected respondent is unavailable to complete the questionnaire.

For example, the first type of noncontact occurs during in-person surveys when an interviewer visits a household unit and finds no one there (but does find clear evidence that the unit is occupied). Noncontacts also occur when contact is made with a household or other sampling unit but the selected respondent is not available to complete the questionnaire at the time of contact. For example, this type of noncontact occurs with in-person surveys when an interviewer visits a sampled address, determines that the address is a household (or other sampled unit), administers the introductory script and respondent selection procedures to someone at the address, and then learns that the selected respondent is not available to complete the interview. This type of noncontact is very similar for telephone surveys and occurs whenever an interviewer dials a case, reaches a household, administers the introductory script and respondent selection procedures for the survey, and learns that the designated respondent is not available at the time of the call. Because contact has been made with someone within the designated sampling unit, cases that result in this type of noncontact usually are considered eligible cases and thus are included when computing survey response rates.

Noncontacts may also occur in mail and Internet surveys, but the nature of these surveys makes it very difficult for researchers to know when this is happening and makes it almost impossible to differentiate between the two types of noncontacts. For example, in a mail survey, the questionnaire may be delivered to a household when the residents are away for the entire field period of the survey. Similarly, in an Internet survey the respondent may be away from email and the Internet for the entire field period, or the questionnaire may be sent to an email address that the respondent does not check during the field period of the survey. Only if the researcher receives information (such as, in the case of an Internet survey, an automated email reply noting that a respondent is away) specifying that the survey questionnaire was sent to and received by the named respondent is the survey researcher able to determine conclusively that a noncontact has taken place.

Because noncontacts usually are considered to be eligible cases or cases of unknown eligibility (depending on the type of noncontact), researchers continue to process these cases throughout the field period. In order to better manage survey sampling pools, many researchers assign different disposition codes to the two different types of noncontacts. These disposition codes allow researchers to manage the sample more precisely. For example, noncontacts in which no contact is made with anyone at the household or other sampling unit often are recontacted on a variety of days and times (or after a specified period of time in a mail or Internet survey) to increase the chances of making contact with someone at the household or other sampling unit. For cases in which contact is made with someone in a household or other sampling unit (but the selected respondent is not available), the researcher can work to identify a good time to recontact the selected respondent. Because these types of noncontacts are a temporary disposition, it is important that researchers learn as much as possible about when to try to contact the selected respondent and then use any information learned to optimize the timing of additional contact attempts and, in doing so, to maximize the chances of converting the noncontact disposition into a completed interview.

Noncontact also can be used as a final disposition if (a) it occurs on the final contact attempt for a case, (b) previous contact was made during the field period but there was no success in completing the questionnaire at that time, and (c) there was never a previous refusal outcome for the case. If there was a previous refusal outcome, the case should be given the final disposition of "refusal" even if the last contact attempt resulted in a "noncontact."

Matthew Courser

See also Busies; Callbacks; Final Dispositions; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Noncontingent Incentives

Noncontingent incentives are traditionally used in survey research as a way of increasing survey response rates. The concept of noncontigent versus contingent incentives is that a noncontigent incentive is given to the respondent regardless of whether the survey is completed, whereas a contingent incentive is given *contingent* on the respondent's cooperation in completing the survey. Typically, the noncontingent incentive would be given at the time the respondent receives the request to complete the survey. This type of incentive is most commonly used with mailed surveys, although it can be used in any survey mode. The most common type of noncontingent incentive in survey research is a monetary incentive paid in the form of cash or as a cash alternative, such as a check. The recent introduction of cash cards and gift cards have made them another viable option for monetary incentive use in surveys. Many nonmonetary incentives have been used to enhance response rates in surveys. Some examples of nonmonetary incentives that can be given as a noncontingent incentive include sweepstakes entries, videos, gas cards, coupons, online credits, small household appliances, books, electronic devices, small gadgets, and knickknacks.

Don Dillman advises that the proper use of noncontingent monetary incentives is one of the most important strategies a researcher can use to improve survey response rates. Social exchange theory postulates that small (i.e., token) noncontingent incentives make the respondent feel socially obligated, that is, "They already gave me something, so now I should do the survey for them."

The scholarly literature shows a clear consensus that the use of a small noncontingent monetary incentive will increase cooperation rates in surveys significantly and is more effective than contingent incentives of considerably greater value. When considering which type of incentive, if any, to use in a particular survey, the researcher should consider the type of survey instrument (mailed, phone, Internet, intercept), the relative importance of the response rate, the level of effort required to complete the survey, the probable motivation of respondents, and the possible need to *differentially incent* members of some hard-to-reach demographic subgroups. For simple and short mailed surveys, short phone interviews, and short Internet surveys, an incentive is not likely to be needed. As the length and complexity of the survey increase or respondent engagement (e.g., level of interest) decreases, the need to consider the use of a noncontingent incentive is likely to increase. Care should be taken to ensure that the incentive offered is appropriate for the respondents being surveyed and does not introduce bias into the behavior of the respondent. An example of an inappropriate incentive would be a free DVD rental offered for participation in a television viewing survey. The respondent behavior that was being measured would most likely be impacted and the results may be biased.

The amount of incentive offered to the respondent should not be out of proportion to the effort required to complete the survey. A respondent who is given "too high" a noncontingent incentive amount as the sole motivating factor in the decision to cooperate in the survey may not answer the survey as accurately as someone else who received a noncontingent incentive of more modest value. Researchers should be aware of this *buying cooperation* phenomenon, which may cause some respondents to provide answers they think the researcher wants from them rather than providing accurate answers. Conversely, some respondents may have become so accustomed to receiving a noncontingent incentive when sampled for a survey that they may dismiss any survey request that does not offer one.

Norm Trussell

See also Contingent Incentives; Economic Exchange Theory; Incentives; Social Exchange Theory

Further Readings

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, *17*, 225–249.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.

James, J. M., & Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality on mail surveys. *Public Opinion Quarterly*, *54*, 346–361.

Singer, E., Van Hoewyk, J., Gebler, N., Trivellore, T., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, *15*, 231–250.

Singer, E., Van Hoewyk, J., & Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, *64*, 189–205.

Trussell, N., & Lavrakas, P. J. (2004). The influence of incremental increases in token cash incentives on mail survey response: Is there an optimal amount? *Public Opinion Quarterly*, *68*, 349–367.

Noncooperation Rate

Noncooperation occurs when a research unit is able to cooperate but clearly demonstrates that it will *not* take required steps to complete the research process. The noncooperation rate compares the number of research units that refuse to cooperate to the number of all potentially eligible units. Noncooperation, along with noncontacts, compromises the majority of survey nonresponse. In survey research, noncooperation often takes three basic forms:

1. *Household refusals* refer to the refusals that occur shortly after the household has been requested to participate in the research and before a designated respondent has been selected from within the household. The reasons of household refusals often involve “Not interested,” “Don’t have time,” and “Don’t do surveys.” No comment hang-ups during a telephone interviewer’s introduction of the survey often occur, especially when the interviewer is speaking English and the respondent lives in a non-English-speaking household.

2. *Respondent refusals* refer to a refusal to participate by the designated respondent selected within the household. Respondent refusals might be harder to convert, since the refusal was given by the person who should be interviewed and not by a gatekeeper within the household.

3. *Breakoffs*, which are a form of partial completion, refer to instances in which the respondent does not continue through the major part of the questionnaire. Breakoffs are different from other partial interviews because the proportion of questions completed in other partials may be considered as sufficient response, depending on the definition of response rate used. The standard of being a sufficient partial response is pre-determined by the researchers; the researcher should always provide a clear definition. For example, legitimate partial completions might be defined as cases with 50% to 94% of the questions answered. Breakoffs are considered as noncooperation, similar to refusals, as some of these are respondents merely hang up on the interviewer after questioning has started without saying anything more.

The most common way to calculate noncooperation rate (NCR) is to use the proportion of all cases in which a sample unit refuses to complete an interview out of all potentially eligible cases; this is comparable to a refusal rate. The numerator includes cases in which the household refuses to start the questionnaire, or the identified respondent refuses to start, or the identified respondent refuses to complete the interview. The denominator is the number of all eligible cases, as indicated in the following formula:

$$\begin{aligned} NCR1 = (&Household\ Refusals + Respondent\ Refusals \\ &+ Breakoffs)/(Interviews + Partials \\ &+ Household\ Refusals + Respondent\ Refusals \\ &+ Breakoffs + Noncontacts \\ &+ Other\ Eligible\ Nonresponse \\ &+ Unknown\ Eligibility) \end{aligned}$$

NCR1 is the most conservative approach to calculating the noncooperation rate.

Other versions of the noncooperation rate differ in the composition of the denominator. The following formula (NCR2) includes estimates of the proportion of cases of unknown eligibility that actually are eligible. By estimating such a proportion (i.e., *e*), researchers aim to make a more precise computation of all potentially eligible units. However, the estimation of *e* must be guided by the best available scientific information on what share the eligible cases make among the unknown cases, and one must not select a proportion simply in order to decrease the noncooperation rate. The basis for the estimate must be explicitly stated and detailed.

$$\begin{aligned} NCR2 = (&Household\ Refusals + Respondent\ Refusals \\ &+ Breakoffs)/(Interviews + Partials \\ &+ Household\ Refusals + Respondent\ Refusals \\ &+ Breakoffs + Noncontacts \\ &+ Other\ Nonresponse \\ &+ [e * Unknown\ Eligibility]) \end{aligned}$$

A third type of noncooperation rate calculation (NCR3) discards all cases of unknown eligibility. It means either a special case of NCR2, in which *e* is assumed to be zero (i.e., that there are no eligible cases among the cases of unknown eligibility) or the rare case in which there are no cases of unknown eligibility. This formula generates the maximum nonresponse rate, since the denominator is the smallest among the three computations:

$$\begin{aligned} NCR3 = (&Household\ Refusals + Respondent\ Refusals \\ &+ Breakoffs)/(Interviews + Partials \\ &+ Noncooperation + Noncontacts \\ &+ Other\ Nonresponse) \end{aligned}$$

Weiyu Zhang

See also Completion Rate; Cooperation Rate; Designated Respondent; *e*; Hang-Up During Introduction (HUDI); Household Refusal; Noncooperation; Nonresponse Rates; Partial Completion; Refusal Rate; Respondent Refusal; Response Rates

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). *Survey nonresponse*. New York: Wiley.

NONCOVERAGE

Every scientific survey has a target population that is operationalized by a sampling frame. Ideally, all units in the sampling frame should match those in the target population on a one-to-one basis. In reality, misalignment between the two occurs and is termed *coverage error*. Noncoverage is one of the elements of coverage error arising from the imperfectness of a sampling frame that fails to include some portion of the population. Because these frames cover less than what they should, noncoverage is also termed *undercoverage*. Noncoverage is the most frequently occurring coverage problem, and it may have serious effects because this problem cannot be recognized easily in the given frame. Because the target population is defined with extent and time, the magnitude of noncoverage depends on the maintenance of the frame. Depending on whether the households or people covered by the frame differ from those not covered, noncoverage may introduce biases (coverage error) in survey estimates.

The classic example of noncoverage is the *Literary Digest* poll predicting Alf Landon as the overwhelming winner over the incumbent president, Franklin D. Roosevelt, in the 1936 election. Although it had surveyed 10 million people, their frame was comprised of the *Literary Digest* readers, a list of those with telephone service, and a list of registered automobile owners. Although the general voter population was the target population of the poll, the sampling frame excluded a large proportion of the target population and, more important, an unevenly higher proportion of the middle- and low-income Democratic voters. The general voter population was more likely to differ in their preference of presidential candidate than those who were covered in the frames. Because the sampling frame failed to represent the target population, the poll results favoring Landon's victory were in error.

There are two main sources of noncoverage error in general population surveys. The first is the problem of covering housing units and the second of covering people within housing units. The effect of these sources in noncoverage differs by survey mode. Telephone surveys are discussed more frequently than other surveys with respect to noncoverage error. Between the two noncoverage sources, coverage of people within housing units in telephone surveys has not been found to be as problematic as coverage of housing units. This is because the landline (wired) telephone survey frames are constructed using a directory listing, random-digit dialing, or a combination of the two. No matter which frame is used, telephone surveys cover households that own a telephone and subscribe to a telephone service. Because ownership of a landline telephone is found to be associated with socioeconomic status, it is acknowledged that the telephone is not the ideal mode for surveys in which the socioeconomically disadvantaged population is an important component of the sample. Since 2002, households in the United States with only cellular (mobile) phones have emerged as another noncoverage problem in traditional telephone surveys whose frames are based on landline telephone numbers. Because the cellular phone–only population in the United States (estimated to exceed 20% of adults in 2008) has distinctive characteristics and its proportion in the general population is continuing to grow, the survey research field is making a concerted effort to alleviate this problem.

Face-to-face and mail surveys use area frames with lists of geographical units or list frames based on addressees or other household identifiers. Frames for face-to-face surveys are further developed by enumerating members in those units. The completeness of the latter component has found to be more problematic than housing unit coverage. This is because enumeration requires the asking about specific members of the unit oftentimes before adequate rapport has been established with whomever answers the door. Interviewers' behavior at the door and the information from the responding member affects the completeness of the enumeration. Males, ethnic and racial minorities, and younger people are more subject to within-unit noncoverage than others.

Noncoverage is a major hurdle for Internet and Web surveys of the general population because a substantial proportion does not have Internet access. The coverage of the Internet is uneven across certain demographic and socioeconomic variables, such as age, education, income, and race/ethnicity. This systematic difference adds complexities and errors in Internet surveys arising from noncoverage. This also results in another distinctive problem, in that it is not feasible even to create a reliable frame for general population Internet surveys.

Sunghee Lee

See also Coverage Error; Internet Surveys; Post-Stratification; Random-Digit Dialing (RDD); Sampling Frame; Target Population; Undercoverage; Unit Coverage; Web Survey, Within-Unit Coverage

Further Readings

Groves, R. M. (1989). *Survey costs and survey errors*. New York: Wiley.

NONDIFFERENTIATION

Survey respondents are routinely asked to answer batteries of questions employing the same response scale. For example, in an effort to understand consumer preferences, respondents might be asked to rate several products on a scale of 1 to 5, with 1 being "very poor" to 5 being "very good." *Nondifferentiation* (sometimes called "straight-lining") occurs when respondents fail to differentiate between the items with their answers by giving identical (or nearly identical) responses to all items using the same response scale. That is, some respondents might give a rating of 2 to all products, producing nondifferentiated answers.

In the survey literature, nondifferentiation is identified as a very strong form of satisficing. According to the notion of satisficing, when respondents are unable to or unwilling to carefully go through all the cognitive steps required in answering survey questions, they may satisfice by looking for an easy strategy or cues to provide a satisfactory (but not optimal) answer. Nondifferentiation is such an easy response strategy that it saves cognitive effort; respondents presumably do *not* retrieve information from memory and do *not* integrate retrieved information into a judgment (or estimation). Instead, they may interpret each question within a battery superficially and select a reasonable point on the response scale and stick with that point for all items in the battery. The answers are thus selected without referring to any internal psychological cues relevant to the specific attitude, belief, or event of interest.

Like other satisficing behaviors, nondifferentiation is most likely to occur when (a) respondents do not have the ability to answer optimally, (b) respondents are not motivated to answer carefully, and/or (c) the questions are difficult to answer. Studies have demonstrated empirically that nondifferentiation is more common among respondents with lower levels of cognitive capacity (such as respondents with less education or with less verbal ability) and more prevalent toward the end of a questionnaire. In addition, nondifferentiation is more prevalent among respondents for whom the question's topic is more personally important.

Nondifferentiation may occur regardless of the mode of data collection. However, there is evidence suggesting that nondifferentiation is more likely to occur with modes that do not promote respondent motivation or use more difficult response tasks. For instance, Web surveys have been shown to promote nondifferentiating responses, especially when questions are displayed in a grid format (i.e., a tabular format where question stems are displayed in the left-most column and response options are shown along the top row). In addition, Web surveys appear to lead to more nondifferentiation than interviewer-administered modes. Within interviewer-administered modes, respondents are found to give more nondifferentiating responses to the telephone surveys than to the face-to-face interviews.

Nondifferentiation is a form of measurement error and thus decreases data quality (both validity and reliability). Of considerable concern, the presence of nondifferentiating responses artificially inflates intercorrelations among the items within the battery and thus suppresses true differences between the items. Therefore, measures should be taken to reduce the extent of nondifferentiation in a survey. Survey researchers, for example, should take measures to help increase respondent motivation to provide thoughtful answers (e.g., interviewers instructing or encouraging respondents to think carefully before answering a survey question) or to lessen the task difficulty (e.g., avoiding a grid format in a Web survey and avoid placing a battery of similar items toward the end of a survey) in order to reduce the extent of nondifferentiation in a survey.

Ting Yan

See also Cognitive Aspects of Survey Methodology (CASM); Measurement Error; Respondent Burden, Respondent Fatigue; Respondent-Related Error; Response Bias; Retrieval; Satisficing; Web Surveys

Further Readings

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of Web and telephone surveys. *Public Opinion Quarterly*, *69*, 370–392.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondents satisficing and social desirability responses bias. *Public Opinion Quarterly*, *67*, 79–125.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

NONDIRECTIVE PROBING

Probing inadequate survey answers for the additional information that may be necessary to fully meet a question's goal(s) is an important element of standardized survey interviewing. In training interviewers to probe effectively, an important distinction should be drawn between nondirective and directive forms of this technique. Unlike directive probing, nondirective probing is designed to encourage and motivate respondents to provide clarifying information without influencing their answers. That is, this approach is specifically designed to be neutral in order to avoid increasing the probability that any specific type of answer is encouraged, or discouraged, from respondents. When nondirective probing is employed, an answer is never suggested by the interviewer. Some examples of nondirective probing of closed-ended questions include slowly repeating the original question or repeating the full set of response options (e.g., "Is that a 'Yes' or a 'No'?"). When asking open-ended questions, some nondirective probe examples include repeating respondent answers, using neutral statements such as, "Could you tell me a little bit more about that?" "I'm not sure I understand what you mean here," "So why do you feel that way?" and "Is there anything else you wanted to say about this?" or simply pausing while respondents collect their thoughts.

Nondirective probing is also important when requesting numerical information. Useful strategies when probing answers to these types of questions include asking respondents to provide more exact information (e.g., "I need a more precise figure if possible"), asking them to select a single number from a range of values initially reported (e.g., "Would you say 2 or would you say 3?"), and asking them to perform any necessary calculations when they provide information using a format other than what was requested (e.g., question: "How old are you?"; answer: "I was born in 1955"; probe: "So how old would that make you?").

In contrast, directive probes are not neutral. They may inadvertently bias respondent answers by limiting the potential range of responses available or by suggesting that some answers are more preferable than others. In probing a closed-ended question, an example of a directive probe would be presenting a truncated range of response options (e.g., answer: "My health is on the low side"; probe: "So, would you say your health is 'only fair' or 'poor'?"). Interviewers often also will construct directive probes to open-ended questions by attempting to reword a respondent's initial answer (e.g., "In other words, you are opposed to income taxes because they are a disincentive to work?"). Similarly, an example of a directive probe to a numeric question might be, "So that means you were 12 when you first smoked a cigarette?". These latter two examples highlight the fact that directive probes can often be answered with a "Yes" or "No" answer.

Although an important element of standardized interviewing, nondirective probes themselves ironically can be only partially standardized and hence are both employed and worded to some extent at the discretion of the interviewer. This variability should also be considered a potential source of measurement error, one that is best confronted through careful training of interviewers regarding the critical nature of their role in conducting standardized survey interviews, as well as the specific goals of each question included in the survey instrument.

Timothy Johnson

See also Bias; Closed-Ended Question; Interviewer Monitoring; Interviewer Neutrality; Interviewer Training; Measurement Error; Open-Ended Question; Probing; Standardized Survey Interviewing

Further Readings

Fowler, F. J. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Gwartney, P. A. (2007). *The telephone interviewer's handbook: How to conduct standardized conversations*. New York: Jossey-Bass/Wiley.

Nonignorable Nonresponse

When patterns of nonresponse (either unit or item nonresponse) are significantly correlated with variables of interest in a survey, then the nonresponse contributes to biased estimates of those variables and is considered *nonignorable*. Recent trends of increasing survey nonresponse rates make the question whether nonresponse is ignorable or not more salient to more researchers.

Since data are only observed for responders, researchers often use participating sample members or members for whom there are complete responses to make inferences about a more general population. For example, a researcher estimating the average income of single parents might use income data observed for single-parent responders to make generalizations about average income for all single parents, including those who did not participate or who refused to answer the relevant questions. The underlying assumption is that single-parent sample members who do not respond or respond with incomplete data are similar to single-parent sample members who participate fully. This implies that the units with missing data or incomplete data are a random subsample of the original sample and do not differ from the population at large.

If this assumption is spurious (i.e., it is *not* true)—that is, units with missing or incomplete data are different in meaningful (nonignorable) ways from the rest of the sample on key variables of interest—then inferences with missing data can lead to biased estimates. For example, if lower-earning single parents have high unit nonresponse rates because they are more difficult to locate and contact, then the estimate of income, the key variable, will be upwardly biased. Thus, when survey participation rates are correlated with key variables, unit nonresponse is likely to be nonignorable.

Essentially every survey has some nonresponse either because of an inability to locate or contact a sample member, or because of a sample member's refusal to participate or to answer certain questions. When researchers make inferences from their sample to the population, then survey response rates are considered an indicator of the representativeness of the data, making the response rate an important criterion of data quality. Because of this, declining response rates make the question of whether or to what extent the nonresponse is ignorable especially important.

The growing problem of nonresponse has led researchers to increase efforts to reduce nonresponse and measure possible nonresponse error. Nonresponse due to noncontact is usually dealt with by improving tracking and locating efforts and by increasing the number of contact attempts at different times of day and days of week to maximize the probability of contact. Survey organizations may provide interviewer training in avoiding or converting refusals. Incentives are used to increase contact rates and decrease refusal rates. Efforts to maintain contact with sample members are used between waves in longitudinal studies to minimize sample attrition. Where nonresponse is due to a physical or mental limitation of the sample member, proxy interviews (e.g., by a family member) may provide key data. In some rare instances, researchers are able to compare survey responses to administrative data in order to measure the impact of nonresponse. Finally, researchers will also make statistical adjustments using external benchmarks such as census data to estimate the impact of nonresponse on their estimates.

Since these can be costly enterprises, they may be inefficient if nonresponse is in fact ignorable—that is, the measures (in the aggregate) that are missing from nonresponders are not different enough from the measures (in the aggregate) taken from responders to change the survey statistics in any appreciable (meaningful) way. Thus there is great interest in separating nonresponse into its components in order to focus on the largest parts of nonresponse that contribute to bias. Targeting resources at a particular component of nonresponse can help minimize bias if the researcher expects the cause of this component of nonresponse to be correlated with key variables.

Danna Basson

See also Ignorable Nonresponse; Missing Data; Nonresponse Error; Nonresponse Rates

Further Readings

Curtin, R. (2005). Changes in telephone survey nonresponse over the past quarter-century. *Public Opinion Quarterly*, *69*, 87–98.

de Leeuw, E., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*, 153–176.

Groves, R. M. (2006). Nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Journal of Official Statistics: http://www.jos.nu

Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 637–645.

Nonprobability Sampling

Sampling involves the selection of a portion of the finite population being studied. Nonprobability sampling does not attempt to select a random sample from the population of interest. Rather, subjective methods are used to decide which elements are included in the sample. In contrast, in probability sampling, each element in the population has a known nonzero chance of being selected through the use of a random selection procedure. The use of a random selection procedure such as simple random sampling makes it possible to use design-based estimation of population means, proportions, totals, and ratios. Standard errors can also be calculated from a probability sample.

Why would one consider using nonprobability sampling? In some situations, the population may not be well defined. In other situations, there may not be great interest in drawing inferences from the sample to the population. Probably the most common reason for using nonprobability sampling is that it is less expensive than probability sampling and can often be implemented more quickly.

Nonprobability sampling is often divided into three primary categories: (1) quota sampling, (2) purposive sampling, and (3) convenience sampling. Weighting and drawing inferences from nonprobability samples require somewhat different procedures than for probability sampling; advances in technology have influenced some newer approaches to nonprobability sampling.

Quota Sampling

Quota sampling has some similarities to stratified sampling. The basic idea of quota sampling is to set a target number of completed interviews with specific subgroups of the population of interest. Ideally, the target size of the subgroups is based on known information about the target population (such as census data). The sampling procedure then proceeds using a nonrandom selection mechanism until the desired number of completed interviews is obtained for each subgroup. A common example is to set 50% of the interviews with males and 50% with females in a random-digit dialing telephone interview survey. A sample of telephone numbers is released to the interviewers for calling. At the start of the survey field period, one adult is randomly selected from a sample household. It is generally more difficult to obtain interviews with males. So, for example, if the total desired number of interviews is 1,000 (500 males and 500 females), and the researcher is often able to obtain 500 female interviews before obtaining 500 males interviews, then no further interviews would be conducted with females and only males would be selected and interviewed from then on, until the target of 500 males is reached. Females in those latter sample households would have a zero probability of selection. Also, because the 500 female interviews were most likely obtained at earlier call attempts, before the sample telephone numbers were thoroughly worked by the interviewers, females living in harder-to-reach households are less likely to be included in the sample of 500 females.

Quotas are often based on more than one characteristic. For example, a quota sample might have interviewer-assigned quotas for age by gender and by employment status categories. For a given sample household, the interviewer might ask for the rarest group first, and if a member of that group were present in the household, that individual would be interviewed. If a member of the rarest group were not present in the household, then an individual in one of the other rare groups would be selected. Once the quotas for the rare groups are filled, the interviewer would start to fill the quotas for the more common groups.

Quota sampling is sometimes used in conjunction with area probability sampling of households. Area probability sampling techniques are used to select primary sampling units and segments. For each sample segment (e.g., city block) the interviewer is instructed to start at a corner of the segment and proceed around the segment contacting housing units until a specific number of interviews are completed in the segment.

In another example, one might select an area probability sample of housing units using multi-stage sampling. At the segment level, the interviewers would be supplied with quotas for adults, assuming one adult is interviewed in each household. The instructions might consist of something simple as alternating between interviewing available males and females in the households they make contact with. In random-digit dialing, a probability sample of telephone numbers can be drawn and a quota sampling method can be used to select one adult from each sample household. In telephone surveys conducted under tight time constraints, the selection of a male or female adult from the household can be limited to adults who are at home at the time the interviewer calls. This eliminates the need for callbacks.

The most famous limitation of this type of quota sampling approach is the failure of the major pre-elections polls, using quota sampling, to accurately predict the results of the 1948 presidential election. The field interviewers were given quotas (with estimates based on 1940 census figures) to fill based on characteristics such as age, gender, race, degree of urbanicity, and socioeconomic status. In addition to the inaccurate quotas, the interviewers were then free to fill the quotas without any probability sampling mechanism in place. This subjective selection method resulted in a tendency for Republicans being more likely to be interviewed within the quota groups than Democrats. The sample thus contained too many Republicans, causing the pre-election polls to incorrectly predict Thomas E. Dewey (the Republican candidate) as the winner.

A major problem with quota sampling is the introduction of unknown sampling biases into the survey estimates. In the case of the 1948 presidential election, the sampling bias was associated with too many Republicans being selected. Another problem with quota sampling is that the sampling procedure often results in a lower response rate than would be achieved in a probability sample. Most quota samples stop attempting to complete interviews with active sample households once the quotas have been met. If a large amount of sample is active at the time the quotas are closed, then the response rate will be very low.

Purposive Sampling

Purposive sampling is also referred to as *judgmental sampling* or *expert sampling*. The main objective of purposive sampling is to produce a sample that can be considered "representative" of the population. The term *representative* has many different meanings, along the lines of the sample having the same distribution of the population on some key demographic characteristic, but it does not seem to have any agreed-upon statistical meaning. The selection of a purposive sample is often accomplished by applying expert knowledge of the population to select in a nonrandom manner a sample of elements that represents a cross-section of the population. For example, one might select a sample of small businesses in the United States that represent a cross-section of small businesses in the nation. With expert knowledge of the population, one would first decide which characteristics are important to be represented in the sample. Once this is established, a sample of businesses is identified that meet the various characteristics that are viewed as being most important. This might involve selecting large (1,000 + employees), medium (100–999 employees), and small (< 100 employees) businesses.

Another example of purposive sampling is the selection of a sample of jails from which prisoner participants will be sampled. This is referred to as two-stage sampling, but the first-stage units are not selected using probability sampling techniques. Rather, the first-stage units are selected to represent key prisoner dimensions (e.g., age and race), with expert subject matter judgment being used to select the specific jails that are included in the study. The opposite approach can also be used: First-stage units are selected using probability sampling, and then, within the selected first-stage, expert judgment is employed to select the elements from which data will be collected. "Site" studies or evaluation studies will often use one of these two approaches. Generally, there is not interest in drawing inferences to some larger population or to make national estimates, say, for all prisoners in U.S. jails.

A clear limitation of purposive sampling is that another expert likely would come up with a different sample when identifying important characteristics and picking typical elements to be in the sample. Given the subjectivity of the selection mechanism, purposive sampling is generally considered most appropriate for the selection of small samples often from a limited geographic area or from a restricted population definition, where inference to the population is not the highest priority. Clearly, the knowledge and experience of the expert making the selections is a key aspect of the "success" of the resulting sample, but it would be difficult to quantify that characteristic of a sample.

Convenience Sampling

Convenience sampling differs from purposive sampling in that expert judgment is not used to select a representative sample of elements. Rather, the primary selection criterion relates to the ease of obtaining a sample. Ease of obtaining the sample relates to the cost of locating elements of the population, the geographic distribution of the sample, and obtaining the interview data from the selected elements. Examples of convenience samples include *mall intercept interviewing*, unsystematically recruiting individuals to participate in the study (e.g., what is done for many psychology studies that use readily available undergraduates), visiting a sample of business establishments that are close to the data collection organization, seeking the participation of individuals visiting a Web site to participate in a survey, and including a brief questionnaire in a coupon mailing. In convenience sampling, the representativeness of the sample is generally less of a concern compared to purposive sampling.

For example, in the case of surveying those attending the Super Bowl using a convenience sample, a researcher may want data collected quickly, using a low-cost method that does not involve scientific sampling. The researcher sends out several data collection staff members to interview people at the stadium on the day of the game. The interviewers may, for example, carry clipboards with a questionnaire they may administer to people they stop outside the stadium an hour before the game starts or give it to people to have them fill it out for themselves. This variation of taking a convenience sample does not allow the researcher (or the client) to have a clear sense of what target population is being represented by the sample. Although convenience samples are not scientific samples, they do on occasion have value to researchers and clients who recognize their considerable limitations—for example, providing some quick exploration of a hypothesis that the researcher may eventually plan to test using some form of probability sampling. On the other hand, some researchers naively treat such samples as equivalent to simple random samples and calculate standard errors based on simple random sampling. Doing this does not produce valid statistical information.

Weighting and Drawing Inferences From Nonprobability Samples

One issue that arises with all probability samples and for many nonprobability samples is the estimation procedures, specifically those used to draw inferences from the sample to the population. Many surveys produce estimates that are proportions or percentages (e.g., the percentage of adults who do not exercise at all), and weighting methods used to assign a final weight to each completed interview are generally given considerable thought and planning. For probability sampling, the first step in the weight calculation process is the development of a base sampling weight. The base sampling weight equals the reciprocal of the selection probability of a sampling unit. The calculation of the base sampling weight is then often followed by weighting adjustments related to nonresponse and noncoverage. Finally, post-stratification or raking is used to adjust the final weights so that the sample is in alignment with the population for key demographic and socioeconomic characteristics. In nonprobability sampling, the calculation of a base sampling weight has no meaning, because there are no known probabilities of selection. One could essentially view each sampling unit as having a base sampling weight of one.

Sometimes nonresponse and noncoverage weights are developed for nonprobability samples, but the most common technique is to use a weighting procedure such as post-stratification or raking to align the nonprobability sample with the population that one would ideally like to draw inferences about. The post-stratification variables are generally limited to demographic and socioeconomic characteristics. One limitation of this approach is that the variables available for weighting may not include key characteristics related to the nonprobability sampling mechanism that was employed to select the sampling units. The results of weighting nonprobability samples have been mixed in the situation when benchmarks are available for a key survey outcome measure (e.g., the outcome of an election).

Recent Developments in Nonprobability Sampling

Finally, it should be mentioned that newer versions of nonprobability sampling have appeared in recent years, some driven by changes in technology. These have generally been labeled *model-based sampling approaches*, and to some degree the use of the term *model-based sampling* has replaced the term *nonprobability sampling*. Web surveys are one new example of nonprobability sampling. Web surveys are generally convenience samples of households or adults recruited to participate in surveys delivered over the Web. The samples are usually set up as panels, that is, a recruited household or adult is asked to respond to some number of surveys over their tenure in the sample. At the recruitment phase, characteristics of the households or adults can be collected. This makes it possible to limit future surveys to households or adults with a specific characteristic (e.g., persons ages 18–24 years, female executives, retirees). Because respondents use the Web to complete the questionnaire, these nonprobability sample Web surveys can often be conducted much more quickly and far less expensively than probability samples.

Another new type of nonprobability sampling is based on selecting email addresses from companies that compile email addresses that appear to be associated with individuals living in households. Some companies set up email panel samples through a recruitment process and allow clients to send a questionnaire to a sample of email addresses in their panel. For both email and Web panel samples, the estimation methods used to attempt to draw inferences from the sample to the population are a very important consideration. The use of propensity scores and post-stratification or raking has been explored by some researchers. The calculation of standard errors, as with all nonprobability samples, is problematic and must rely on model-based assumptions.

Another relatively new nonprobability sampling method is known as *respondent-driven sampling*. Respondent-driven sampling is described as a form of snowball sampling. Snowball sampling relies on referrals from an initial nonprobability or probability sample of respondents to nominate additional respondents. It differs from multiplicity sampling in that no attempt is made to determine the probability of selection of each subject in the target population. Snowball samples are sometimes used to select samples of members of a social network in the situation when no complete list of such members exists and the costs of doing a probability sample would be prohibitive. Respondent-driven sampling has most often been employed for surveys of very rare populations in relatively small geographic areas, such as a city or county.

The use of probability sampling, as championed by Leslie Kish and other important statisticians, has resulted in probability sampling being employed in most surveys conducted by the U.S. government. For commercial research, probability sampling methods and nonprobability sampling methods have been employed. More recently, as the cost of collecting data has risen, a considerable amount of the commercial research conducted in the United States has moved to nonprobability sampling methods. For surveys conducted for the federal government, model-based sampling methods have been used in some situations. During the coming years, it is possible that the use of probability sampling will decline further. It is therefore important that more research be conducted to further assess biases from using nonprobability samples and devise strategies to both measure and adjust for these biases.

Michael P. Battaglia

See also Area Probability Sample; Convenience Sampling; Design-Based Estimation; Kish, Leslie; Mall Intercept Survey; Model-Based Estimation; Multiplicity Sampling; Multi-Stage Sample; Probability Sample; Propensity Scores; Purposive Sample; Quota Sample; Raking; Random Sampling; Respondent-Driven Sampling (RDS); Sample; Sampling; Simple Random Sample; Snowball Sampling; Stratified Sampling; Weighting; Web Survey

Further Readings

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Nonresidential

Nonresidential dispositions occur in telephone and in-person surveys of the general public when a case contacted or called by an interviewer turns out to be a business or other type of nonresidential location such as a hospital, government office, or library. The nonresidential disposition also usually includes institutions such as prisons, sanitariums, work camps, and group quarters such as military barracks and college dormitories. Nonresidential cases are considered ineligible for a survey of the general public survey because conducting an interview with these cases would violate critical assumptions of probability sampling. Although the proportion of nonresidential cases in a telephone or in-person sample varies based on the sampling area, the nonresidential survey disposition tends to be fairly common, with nonresidential numbers comprising up to a significant minority of the telephone numbers in a landline telephone random-digit dial (RDD) sample. More recent technologies that are capable of screening out numbers in an RDD sample without interviewers calling the number have substantially reduced the proportion of nonresidential numbers in many telephone survey samples.

One special challenge in telephone surveys is call forwarding technology, which allows one telephone number to be transferred to another number. For example, if a residential telephone number is transferred to another telephone number within the same household, the case may still remain eligible if the survey uses weighting techniques to adjust for unequal probabilities of selection (since the residential unit contacted effectively has an additional telephone line, and thus, a greater chance of being sampled). If a nonresidential telephone number is forwarded to a residence (e.g., a business number being forwarded to the owner's home residence), the case should be considered ineligible. More detailed rules may be needed for special (but uncommon) cases in which the number for a residence outside the sampling area is forwarded to another residence inside the sampling area. Finally, a residence and a business occasionally may share the same telephone number; these cases should be treated as eligible for the interview. Only numbers that ring exclusively to a business or other nonresidential unit should be given the nonresidential disposition.

For in-person surveys, two special cases can make it more difficult to determine whether a case is truly nonresidential. For example, although the primary unit at an address might be a business or an institution, it is important to ensure that there is not a residential housing unit within the larger unit (such as an apartment above a business or a warden's house on the grounds of a prison). An additional challenge is posed by vacation and other seasonal homes, and special rules may need to be developed to properly determine the disposition of these housing units.

The challenges and special cases mentioned illustrate that at times it may be difficult to determine whether a telephone number or housing unit is residential. Although these instances are fairly uncommon, it important to ensure that there is definitive evidence that the case is nonresidential before applying the nonresidential disposition to a case. Obtaining this evidence may require additional investigation (such as talking to neighbors or documenting visible signs that the unit is uninhabited or not used as a residence). In surveys of the general public, a nonresidential outcome is treated as final and ineligible and thus is not used in response rate calculations.

Matthew Courser

See also Call Forwarding; Dispositions; Final Dispositions; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Nonresponse

Nonresponse refers to the people or households who are sampled but from whom data are not gathered, or to other elements (e.g., cars coming off an assembly line; books in a library) that are being sampled but for which data are not gathered. The individual nonrespondents on a survey contribute to the *nonresponse rate*, the aggregate tally of how many did not provide data divided by how many "should" have. One minus the nonresponse rate is, of course, the overall response rate.

Classification

Calculating nonresponse is not as simple as it initially seems. One survey may have mainly nonrespondents who could not be reached (i.e., contacted) by the researchers. Another reaches most of the sampled persons, but a large proportion refuse the survey. A third has an easily reached, cooperative set of people in a sampling frame, but many do not possess a characteristic that is part of the definition of the desired survey population. These reasons for nonresponse, usually termed *noncontact, refusal*, and *ineligibility*, respectively, are differentiated within most classification procedures for nonresponse, such as the *Standard Definitions* established by the American Association for Public Opinion Research (AAPOR). The components of nonresponse may differ across the various contact types or modes of survey. For example, the noncontact count may be higher on a telephone survey in which calls to some numbers are never answered than on a postal survey in which a letter mailed is presumed to be a contact, except for a few nondeliverables returned by the post office. On the street, as distinguished from the logical world of survey methodology research articles and textbooks, the distinctions among noncontact, refusal, and ineligibility are not always watertight. Is a foreign-born person who opts out of a survey due to language problems, or an elderly citizen with impaired hearing or eyesight, an ineligible or a refusal? Only the person himself or herself really knows. What of those who, in a face-to-face household survey, recognize the approaching person as a survey taker and decline to answer the doorbell? Refusal or noncontact?

Predictability

Nonresponse, like many social behaviors, is only weakly predictable at the level of the individual yet becomes more ordered at the aggregate. Any experienced survey methodologist, told the type of population to be sampled, the mode of contact (i.e., telephone, in-person, mail, Web, or multi-mode), the length and content of the questions, the resources available for callbacks and follow-ups, and whether or not payments or other incentives are to be used, can give an informed estimate of the final response and nonresponse rates. Information on the sponsor, and whether the fieldwork is to be conducted by a government agency, a survey unit at a university, or a commercial firm, will add further precision to the estimates. At the same time, and in most cases, the attempt at prediction of whether one person versus another will respond or not generates only slight probabilities. Always depending on the particulars of the survey, people of greater or lesser education, male or female, racial majority or minority, young or old, may be disproportionately among the nonrespondents. Such effects often appear in surveys, but few generalities are possible. Perhaps the safest is the tendency for middle, rather than upper or lower, socioeconomic status people to respond: the *middle-class bias* sometimes attributed to surveying. Even this, however, is not a certainty. For example, on a survey whose results will help generate funds for social programs, the lower socioeconomic strata may have good reason to respond. If the survey is conducted by an agency of government, there may be some deference to the authority of the sponsor among the lower strata, also contributing to probability of response. Offsetting those advantages might be a subcultural hesitancy and anxiety about participating in surveys. The probability of an individual responding to a survey remains inherently hard to predict, because so many factors enter the mix. (An exception being panel surveys in which response behavior in a previous wave can be highly predictive of subsequent wave participation.)

A Historical Perspective on Survey Nonresponse

It used to be firmly believed by survey methodologists that the response versus nonresponse rates on a survey accurately indicated the quality of the data. One rule of thumb, to which several generations of students were exposed via a widely used textbook by Earl Babbie, was that a mailed survey with 50% response was "adequate," 60% "good," and 70% "very good." These notions were advanced in the early 1970s, when the highest quality surveys still tended to be face-to-face and the telephone survey was still under development. Face-to-face ("personal") interview surveys of the national population in the United States, if of high quality, were expected in this era to have nonresponse rates of only some 20 to 25%. By the mid-1970s, however, a growing number of voices were detecting erosion in response rates to personal interview surveys. The issue even was noted in *The New York Times* on October 26, 1975. The telephone survey was meanwhile gaining acceptance among survey researchers as they perfected the generation of sampling frames via random-digit dialing (RDD). Research was also under way during the 1970s on enhancing the validity of mailed surveys, led by Don Dillman's comprehensive strategy known as the "total design method" or TDM.

By the end of the 1990s, a second crisis in survey response rates had developed. This time, the center of attention was the telephone survey, for face-to-face household interview surveys had become rare due to their astronomical cost when fielded to a national population. Telephone surveys by commercial firms, when dealing with topics not deemed highly salient by the public, were now highly unlikely to breach the 50% response mark, the guideline used some 30 years earlier to define adequate quality in a survey. Further, as the proportion of U.S. households with cell phone only (no landline) approached 10% in 2005, problems of sample coverage on telephone surveys were mounting.

Declining Response Rates and Implications

As one report of declining responses followed another, as the pattern was detected not just in the United States and Canada, but throughout Europe as well, researchers began to reevaluate the consequences of high nonresponse for the validity of a survey. Studies appeared with examples in which nonresponse was *ignorable*, that is, did not seem to seriously bias results from the survey. To be sure, these results did not mean that nonresponse no longer mattered, but they did challenge the simple rule of thumb that, if a survey were to be trusted, a majority of those who could respond should respond. By 2008, it is more clearly realized that one survey could have low, but seriously biasing nonresponse, another much higher but more ignorable nonresponse. A survey, for example, on a rare deviant behavior might be nonthreatening for abstainers but precipitate heavy nonresponse among practitioners. If the population sampled were somewhat "captive," such as students in a school, the response rate could be high, but the nonrespondents who refused to sign permission slips might be precisely the respondents most urgently needed for the survey estimates to be accurate. Nonresponse would, in contrast, prove ignorable if nonrespondents were a near-random subset of the sample, even if the proportion of nonrespondents to respondents were very high.

Although the notion of ignorable nonresponse has provided some reassurance about the continuing usefulness of the survey method for evidence-based social policy, academic studies, and market research, rising nonresponse rates still require attention. When there are nonrespondents in a survey, the researcher can never be certain that bias does not exist for at least some variables.

Current Theories

Survey methodologists thus put considerable resources into the quality of survey fieldwork. Since part of the problem of high survey nonresponse in the 21st century is lifestyle based—for example, the diminished predictability of when a householder will be reachable by telephone at home—the number of repeat calls or "callbacks" has had to be increased for high-quality telephone surveys. Survey research organizations are also now more involved in attempts at "conversion" of initial refusers than they may once have been. Considerable experimentation has been conducted with incentives for surveys. Modest gifts or small sums of money work especially well on the self-enumerated modes of survey, such as mailed or Web contact. The incentive can be delivered up-front, "pre-paid," not conditional on the response taking place. Somewhat counterintuitively, pre-paid rewards prove more effective than the "post-paid" rewards given only after the survey is successfully collected from the respondent. Naturally all such techniques add to the costs of surveying and tend to slow down the speed at which data can be collected.

Especially for commercial survey firms, who must watch the balance sheet, it is difficult to maintain response rates near historical norms. To be sure, surveys via the Internet—the *Web survey*—hold some promise for controlling the spiraling costs of survey research. Since the year 2000, growing attention to the Web mode has been evident in the research journals. At the moment, Web surveys work best for well-defined subpopulations having known email addresses. University students, for example, are routinely being surveyed by university administrators and academic researchers, with response rates in the 40s when carefully conducted. NSSE, the National Survey of Student Engagement, is one well-known example. The general public is harder to survey adequately via the Web. A survey questionnaire simply dumped onto an Internet site for those Web passers-by who wish to answer (sometimes as many times as they like!) is not a serious option. Better-quality Web surveys mimic the traditional mailed survey. A sampling frame exists, consisting in the Web case of email addresses, and a sequence of contacts is carried out. Provision is made via passwords to prevent one person answering several times. It is imaginable that convergence will take place between the methods of the mailed survey and telephone RDD. That is, if plausible, randomly generated email addresses were created, with a filtering system to capture the working ones, it would be possible to field Web surveys of general populations. Technical issues would arise around defining exactly what this population was, for the samples would sprawl across national boundaries. The method might be no more uncertain, however, than the current interest in creating "panels" of reusable respondents.

If one answer to declining response rates is thus technological innovation, another draws from the traditional conceptual tools of the social sciences. The growing resistance to surveys, especially those by telephone, has prompted renewed theorizing about what kind of behavior is involved when people do or do not respond to a survey. Many theoretical approaches appear within the research literature, but two main approaches are especially prominent. *Social exchange theory*, that dates to the mid-1970s and as applied to survey nonresponse, asks if response decisions are best understood as reasoned actions reached after a mental processing of relative costs and benefits. The cognitive heuristics people use in decision making received increasing attention toward the end of the 1980s. Here response behavior is viewed as somewhat impulsive action largely bypassing the consciousness.

Along with decision-making styles, survey methodologists theorize about the role of cultural factors. Do people, for example, hold attitudes toward the survey method in general, accepting or declining specific surveys largely on the basis of the attitude, or is the behavior "situational," mainly determined by what else is going on when a respondent is contacted? Research on attitudes toward the survey method has intensified in the new century. Such attitudes do tend to relate with survey response behavior, but the ties are weak, making it an overstatement to say that the respondents on any particular survey are only those in favor of the method. On the other hand, when it does occur, this is highly nonignorable, biasing nonresponse. The main determinant of who does and does not participate in a survey seems to lie within the multitude of contingent, situational circumstances. People may be busy or idle when the telephone rings; the fieldworker may succeed or fail in striking an instant rapport with the prospective interviewee. These are closer to random events, which more likely result in less-biasing nonresponse.

Complicating these attempts at theoretical interpretation of survey nonresponse is the heterogeneity of human populations. Not all topics hold the same interest for all people—quite the opposite. Research by Robert M. Groves and colleagues shows that response rates can exceed the norm by tens of percentage points when the topic of the survey and the subpopulation being sampled match up well. It follows from this "leverage-saliency" interpretation of nonresponse that techniques such as cash incentives will be proportionately most effective when the topic–population match, thus the salience of the topic, is weak. As such research accumulates, it is increasingly realized that just knowing how many of the eligible people in a sample responded to a survey is insufficient information. People assessing the validity of a survey need to know why people responded and why others did not.

The study of survey nonresponse is thus both a technical issue for methodologists and a rich conceptual puzzle for social science. Ideas about survey nonresponse adapt to the tectonics of social change, making this research topic a living, changing body of ideas rather than a cemetery of textbook certainties.

John Goyder

See also American Association for Public Opinion Research (AAPOR); Leverage-Saliency Theory; Nonresponse Error; Nonresponse Rates; Panel; Random-Digit Dialing (RDD); Social Exchange Theory; Standard Definitions; Total Design Method (TDM)

Further Readings

de Leeuw, E., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41–54). New York: Wiley.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method for surveys*. New York: Wiley.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: Wiley.

Goyder, J., Boyer, L., & Martinelli, G. (2006). Integrating exchange and heuristic theories of survey nonresponse. *Bulletin de Méthodologie Sociologique*, *92*, 28–44.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675.

Groves, R., Singer, E., & Corning, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, *64*, 299–308.

Stoop, A. L. (2005). *The hunt for the last respondent: Nonresponse on sample surveys*. The Hague, Netherlands: Social and Cultural Planning Office.

Nonresponse Bias

Nonresponse bias occurs when sampled elements from which data are gathered are different on the measured variables in nonnegligible ways from those that are sampled but from which data are not gathered. Essentially, all surveys are likely to have some degree of nonresponse bias, but in many cases it occurs at a very small and thus a negligible (i.e., ignorable) level. The size of the bias is a function of (a) the magnitude of the difference between respondents and nonrespondents and (b) the proportion of all sampled elements that are nonrespondents. Thus, even if only one of these factors is large, the nonresponse bias may well be nonnegligible.

There are three basic types of survey nonresponse. The first is refusals, which occur when sampled individuals or households decline to participate. The second is noncontacts, when sampled individuals are never reached. The third type of nonresponse consists of situations in which the interviewer cannot communicate with the sampled person because of a language barrier or some mental or physical disability. Most nonresponse is the result of refusals and noncontacts.

It has long been thought that response rates are a good indicator of survey quality and nonresponse bias; however, recent research has challenged this notion. This is encouraging news for researchers because it means that surveys with lower response rates are not necessarily more biased than those with higher response rates. But this does not mean that nonignorable nonresponse bias cannot or will not occur. Nonresponse bias will be present when the likelihood of responding is correlated with the variable(s) being measured, and this correlation can vary across variables even within the same survey.

Nonresponse bias has been a growing concern to survey researchers as response rates have declined over the years. There are a variety of reasons for this, including an increase in refusals with the rise of telemarketing and an increase in technologies to screen calls such as answering machines, voicemail, and caller ID. It is important to consider nonresponse due to refusals and noncontacts separately because the characteristics of refusers can be different from those who are difficult to reach.

Researchers can use several methods to maximize response rates; much recent research has focused on the correlation between response rates and nonresponse bias; and strategies for reducing nonresponse bias are being designed.

Maximizing Response Rates

Researchers can try to maximize response rates in various ways. For refusals, interviewers can be trained on refusal avoidance strategies, and refusal conversions can be conducted in an attempt to include the less cooperative in the sample. For noncontacts, repeated contacts (in the case of in-person surveys) and callbacks (in the case of telephone surveys) can be made at different times of day and on different days of the week, so that people who are more difficult to reach are included. And language barriers can be overcome by using bilingual interviewers. Repeated reminder mailings in the case of mail surveys and Internet surveys can also be deployed to reduce nonresponse in those types of surveys. Other techniques have been used to increase response rates, such as sending advance letters and using incentives.

Recent Research

Through the 1990s, a common assumption in the field of survey research was that surveys with higher response rates always were more accurate (i.e., had lower nonresponse bias) than those with lower response rates. But this assumption has been challenged by recent research on opinion questions in telephone polls and exit polls. In telephone polls, Scott Keeter and his colleagues, and separately Richard Curtin and his colleagues, have found that fairly large changes in response rates had a minimal impact on their survey estimates. Using a large number of in-person exit polls over different election cycles, Daniel Merkle and Murray Edelman found little or no relationship between response rates and survey bias.

This research has been encouraging to survey researchers because it shows that surveys with lower response rates are not necessarily less accurate. However, this does not mean that survey researchers do not need to worry about nonresponse bias. Robert Groves conducted an extensive review of previous nonresponse studies, most of which focused on nonopinion variables, and found a number of instances in which nonresponse bias was present, even in some surveys with very high overall response rates. He found that response rates were not a good predictor of such bias, consistent with the previous research mentioned. Groves also found that the magnitude of nonresponse bias even differed across variables *within the same survey*, further pointing out the limitation of using a survey's response rate as a measure of data quality and as an indicator of the presence of nonresponse bias.

One reason that response rates are not good predictors of nonresponse bias is that they tell researchers nothing about the second critical component of what determines the magnitude of bias: the size of the difference on key survey measures between respondents and nonrespondents. For variables on which respondents are not different than nonrespondents, surveys with lower response rates will be as accurate as those with higher response rates. The difficult part is knowing when this is the case, because researchers often do not have data on nonrespondents.

Groves explains this by noting that nonresponse bias is a function of the correlation between the survey variable(s) of interest and peoples' likelihood of participating in the survey, called *response propensity*. When there is no relationship between the survey variable and response propensity, then there will not be any nonresponse bias. Nonresponse bias increases as the correlation between response propensity and the survey variable increases.

For example, if one were conducting a survey in the evenings measuring the frequency of dining out, there would be a negative correlation between the survey variable (frequency of dining out) and the likelihood of responding to the survey. Those who dine out more often would be less likely to be home at the time of the survey contact and therefore would be less likely to participate. In this case, the survey would understate the amount that people dine out, because those who do so more often would be less likely to be available to respond.

Reducing Nonresponse Bias

Nonresponse bias can be reduced by *decreasing the correlation* between response propensity and the survey variable. This could be done in the dining-out example by extending the survey calling times to other parts of the day in addition to evenings. The correlation between the frequency of dining out and response propensity would decrease and so would the magnitude of the nonresponse bias on this variable.

It is also possible for procedures designed to increase response rates to actually increase nonresponse bias. This will occur when the mechanism designed to increase response rates *increases the correlation* between response propensity and the survey variable. For example, consider a survey situation in which people with lower incomes are initially more likely to respond, and the researcher includes a monetary incentive in the survey design as a way to increase the response rate. If the appeal of the monetary incentive is negatively correlated with income, the incentive could increase the response rate but also increase nonresponse bias on variables related to income.

Another important way survey researchers try to decrease nonresponse bias is by applying weighting adjustments to the data, called *post-stratification*. A common approach is to weight survey data to match census demographics such as age, race, gender, and education. These types of adjustments assume that respondents and nonrespondents in a given demographic group are similar on the other survey variables measured. When this is the case, such weighting adjustments will decrease bias on variables that are highly correlated with the weighting variables. Weighting adjustments that can take into account factors that influence response propensity will be more successful at reducing or eliminating bias.

Daniel M. Merkle

See also Callbacks; Ignorable Nonresponse; Noncontacts; Nonignorable Nonresponse; Nonresponse Error; Nonresponse Rate; Post-Stratification; Post-Survey Adjustments; Refusals; Refusal Avoidance Training (RAT); Refusal Conversion; Refusal Rate; Response Propensity; Response Rates

Further Readings

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the Index of Consumer Sentiment. *Public Opinion Quarterly*, *64*, 413–428.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675.

Keeter, S., Miller, C., Kohut, A., Groves, R., & Presser, S. (2000). Consequences of reducing nonresponse in a large national telephone survey. *Public Opinion Quarterly*, *64*, 125–148.

Merkle, D. M., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 243–258). New York: Wiley.

NONRESPONSE ERROR

During the past decade, nonresponse error—which occurs when those units that are sampled but from which data are not gathered differ to a nonignorable extent from those sampled units that do provide data—has become an extremely important topic to survey researchers. The increasing attention given to this part of the total survey error is related to the observation that survey participation is decreasing in all Western societies. Thus, more efforts (and cost expenditures) are needed to obtain acceptably high response rates.

In general, the response rate can be defined as the proportion of eligible sample units for which an interview (or other form of data collection) was completed. The calculation of the standard response rate is straightforward: the number of achieved interviews divided by the number of sample units for which an interview could have been completed. These are eligible sample units: completed interviews, partial interviews, noncontacted but known eligible units, refusals, and other noninterviewed eligible units.

Simple Model

Although the nonresponse rate is important information used to evaluate the nonresponse error, that rate is only one component of the nonresponse error. The biasing effect of nonresponse error is also related to the difference between respondents and nonrespondents. A simple model for a sample mean can be used to illustrate this point.

Given a sample with n units, r of these n units participated in the survey and nr units did not: $n = r + nr$. Y is a metric characteristic, and $\overline{Y}_r =$ mean estimated for the r respondents, $\overline{Y}_n =$ mean estimated for all n sample units, and $\overline{Y}_{nr} =$ mean estimated for the nr nonrespondents. The estimated mean for all the units in the sample ($\overline{Y}_n$) equals to a weighted sum of the estimated mean for the respondents ($\overline{Y}_r$) and the estimated mean for the nonrespondents ($\overline{Y}_{nr}$). The latter is weighted by the nonresponse rate $\left(\frac{nr}{n}\right)$, the former by the response rate $\left(\frac{r}{n}\right)$. This results in the following formal specification:

$$\begin{aligned}\overline{Y}_n &= \frac{r}{n}\overline{Y}_r + \frac{nr}{n}\overline{Y}_{nr}\\ &= \left(1 - \frac{nr}{n}\right)\overline{Y}_r + \frac{nr}{n}\overline{Y}_{nr}\\ &= \overline{Y}_r - \frac{nr}{r}(\overline{Y}_r - \overline{Y}_{nr})\\ \overline{Y}_r &= \overline{Y}_n + \frac{nr}{r}(\overline{Y}_r - \overline{Y}_{nr})\end{aligned}$$

This expression makes it clear that the estimated mean for the respondents is equal to the estimated mean for all units in the sample plus a factor that expresses the biasing effect of the nonresponse error. When there is no nonresponse, then ($nr = 0 \rightarrow r = n$): $\overline{Y}_r = \overline{Y}_n$. In this situation, the estimated mean for the r respondents is equal to the mean estimated for all n sample units. This signifies that there is no nonresponse error. This is also the case when the estimated mean of the respondents and the nonrespondents are equal: $\overline{Y}_r = \overline{Y}_{nr}$. In this case, the decision to participate is uncorrelated with Y, and as far as Y is concerned, there is no difference between the group of respondents and the group of nonrespondents.

Although this nonresponse model for a sample mean is simple, it shows that the reduction of nonresponse error operationally is not straightforward. For example, with an incentive one can increase the response rate, but it is possible that some persons are more susceptible to the incentive than others. When the susceptibility to incentives is related to a substantive relevant characteristic, the use of the incentive will *increase* the difference between respondents and nonrespondents with respect to that characteristic, thus increasing nonresponse error. In this situation, a higher response rate due to the incentive does not result in a smaller nonresponse error; instead, just the opposite occurs.

The model also illustrates that for the evaluation of the nonresponse error one must both compare the respondents with the nonrespondents and calculate the response rate. After defining some additional basic concepts, the information necessary for this evaluation follows.

Basic Concepts

Until now, the sample units in the example given were divided or classified into two general groups:

respondents and nonrespondents. The respondents are eligible sample units with a completed interview; nonrespondents are eligible units that are not interviewed. *Ineligible units* in a sample of individuals include those that are deceased; emigrated or have left the country for a long period of time; units residing in an institution and unoccupied or demolished premises. In a sample of households or addresses, the ineligible units include the following: unoccupied or demolished premises; premises under construction; nonresidential addresses; addresses occupied, but no resident households; and addresses occupied by resident households, but no eligible respondents. For some units there is not enough information; these units are classified as units of *unknown eligibility* (e.g., unknown whether an eligible person is present in an existing housing unit).

In a standard in-person household survey, there are two main reasons for nonresponse: refusal and noncontact. Mostly, interviewers are instructed to contact sampling units according to a prescribed contact procedure. For example, the contact procedure may specify at least four contact attempts, including at least one visit in the evening and one during the weekend, and that these visits should be spread over at least two different weeks. Even then it is not always possible to contact all the units in the sample. These units that are not contacted at the end of the contact procedure receive the final disposition of "noncontact." Refusals are contacted sample units with a negative reaction to the request to participate in a survey. Given these two main reasons for nonresponse, one can further segment the group of nonrespondents into three subclasses: (1) refusals, (2) noncontacts, and (3) other nonrespondents. The concepts just defined can be represented in a simple diagram (see Table 1).

The classification shown in Table 1 is useful to define and calculate some rates that are relevant to evaluating nonresponse error. As already mentioned, the response rate is the total number of respondents divided by the number of eligible sampling units. Among the group classified as "unknown eligibility" (*ue*) one can estimate the proportion of eligible units (*pe*) by using the current survey. This proportion (*pe*) equals $e/(e + ie)$.

Using these definitions, the calculation of the *response rate* is straightforward: $r/(r + rf + nc + on + (pe \times ue))$. One can also calculate the proportion of nonrespondents in the group of eligible units: *nr*/ (*e nonresponse rate* + $(pe \times ue)$). It is a standard practice to calculate the *refusal rate* as $r/(r + rf)$.

Information to Evaluate Nonresponse Error

Information from all the units in the sample (respondents and nonrespondents) is needed to evaluate the nonresponse error in a survey. This is the case both to calculate the rates discussed previously and to compare the group of respondents with the group of nonrespondents. Obtaining information from nonrespondents seems at first a contradiction in terms. Although it is a difficult task to collect information from the nonrespondents, there are some opportunities. Sometimes one can use the sampling frame. The sampling frame is the listing of all units in the target population. Sometimes the sampling frame (e.g., national register, staff register) contains information about some basic sociodemographic characteristics of the units (e.g., age, gender, civil status, educational level, professional group). When this kind of information is available, it offers an excellent opportunity to compare respondents and nonrespondents using some relevant background characteristics. Another extremely useful source of information

Table 1 Classification of final fieldwork outcomes in a survey

Sampling Units (n)			
Eligible (e)		*Unknown Eligible (ue)*	*Ineligible (ie)*
Respondents (*r*)	Nonrespondents (*nr*)		
	- Refusals (*rf*)		
	- Noncontacts (*nc*)		
	- Other nonrespondents (*on*)		

to evaluate the nonresponse error is the *paradata* collected by means of the case control form. A case control form is an essential instrument for monitoring the field work or interviewer activities of a survey. A case control form is filled out by the interviewer; it contains information about all contact attempts and contacts—date, day of the week, time, mode of the visit (personal, personal but only intercom, telephone, etc.)—and outcome of contact (interview, appointment, no contact, refusal, etc.). Also, the reason for refusal mentioned by the respondent (e.g., bad timing, not interested, don't know enough about the topic, too difficult for me) can be registered on the contact form or on a separate refusal report form. Sometimes in-person interviewers are also asked to record on these forms a number of observable area and dwelling characteristics (e.g., type of dwelling, physical state of dwellings in the area), as well as observable characteristics and other information about the person who refuses.

It is important that the case control form is filled out for all the eligible units in the sample and that the information is available for both respondents and nonrespondents. The information from the case control form is essential for calculating the response rates. (One can say that a case control form of some description is a vital part of a well-designed and organized survey.) Without information from the case control form, a profound evaluation of the nonresponse error is really not possible. One must realize that using a case control form makes the requirements of the field work activities in a survey more severe and complex. For example, an increase of the survey cost is a logical consequence.

Reducing Nonresponse Error

Although the structure of the simple model of nonresponse error makes it clear that reduction of nonresponse error is not just a matter of increasing the response rate, much of nonresponse research is related to factors that can have a positive effect on the decision to participate in a survey interview. In a frequently cited model of survey participation developed by Robert M. Groves and Mick P. Couper in the late 1990s, the (doorstep) interaction between the interviewer and the respondent is the only factor with a direct effect on the respondent's decision to cooperate or to refuse. Other factors in the model have an indirect effect through the interaction. These factors can be classified into two categories: (1) factors out of researcher control and (2) factors under researcher control. The factors out of researcher control are related to the general social context in which the survey is organized (political and economic conditions, survey-taking climate, neighborhood characteristics) and to respondent's characteristics (sociodemographic characteristics, knowledge of the topic, experience with surveys, psychological predisposition). Survey design features (respondent selection, survey topic, mode of administration) and interviewer characteristics (sociodemographic characteristics, experience, expectation, ability) are factors that the researcher can control.

Given the central role in the model of the interaction between interviewer and respondent, the interviewers' training to obtain positive reactions to the request to participate is one important factor in increasing survey participation. During the training, interviewers must become familiar with adequate doorstep strategies. Two techniques must be mentioned in this context: *tailoring and maintaining* interaction. In contrast with the idea that interviewer behavior must be standardized, the initial interaction with the respondent is not directed by a standard script. Interviewers must tailor their interaction with the respondent. They must read the cues from the respondents and adapt their interviewing approach, thereby averting a refusal. Some respondents, for example, are more sensitive to the topic of the questionnaire; others are more willing to participate if they believe that others will also participate. During the interaction, the interviewer must play upon the reasons that are important in the respondent's decision process. To obtain information about which reasons will dominate the respondent's decision, a second interaction principle, maintaining interaction, must be applied. Interviewers should try to avoid provoking a quick negative reaction from the respondent. It is important that they try to prolong the interaction as long as possible, so that they get enough information to persuade the respondents adequately. In this way, maintaining interaction is vital to the tailoring principle. Increasing the number of contact attempts can also be considered as a special kind of maintaining interaction.

It is not only important to train the interviewers to achieve contact and to persuade the respondents, but also to closely monitor the interviewer's field work activities (e.g., contact strategies, response rate, contact rate, and refusal rate for each interviewer) so that information can be used to improve fieldwork performance.

It is clear that a few interviewers with exceptionally high refusal and noncontact rates can have a serious impact on the field work results. Detection of these rates at an early stage in the field work is crucial.

In addition to the interviewer's training and follow-up, it is generally accepted that an advance letter (or other form of advance contact) to the sample units has a positive effect on the response rate. In the advance letter, the topic and the intention of the survey interview are introduced, the field work organization and the responsible authority are mentioned, and privacy and confidentiality are emphasized. When incentives are used, one can announce the incentive in the advance letter. Pre-paid (noncontingent) cash incentives in particular can increase the willingness of a sample unit to cooperate.

Refusal conversion, that is, obtaining information from initially reluctant respondents, is another strategy that one can use to increase the response rate. It is typical that an experienced interviewer with good response rates is used to try to make a refusal conversion attempt with any reluctant respondents. This means that a refusal is recontacted by another interviewer and asked again to participate in the survey. Due to privacy reasons, this procedure is illegal in some countries.

A well-organized field work for a survey offers possibilities to minimize nonresponse error during the data collection period and to assess the nonresponse error at the end of the field work. In a well-organized survey, an adequate sampling frame is available; interviewers are well trained to contact and persuade the respondents; an informative and carefully edited advance letter is used; a well-designed case-control form is filled out after each contact attempt with each sample unit; if necessary and possible, a refusal conversion procedure is implemented, and accurate and timely monitoring of the field work is organized. All these characteristics are not only relevant to decrease and to evaluate nonresponse error but also are very useful to improve survey data quality in general.

Geert Loosveldt

See also Advance Contact; Advance Letter; Control Sheet; Error of Nonobservation; Ignorable Nonresponse; Incentives; Ineligible; Interviewer Monitoring; Interviewer Training; Noncontact; Nonignorable Nonresponse; Nonresponse Rates; Paradata; Refusal; Refusal Avoidance Training (RAT); Refusal Conversion; Refusal Report Form (RRF); Response Rates; Sampling Frame; Standard Definitions; Tailoring; Total Survey Error (TSE); Unknown Eligibility

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. New York: Wiley.

Groves, R. M. (2006). Nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, R. M., & Couper, M. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2002). *Survey nonresponse*. New York: Wiley.

Nonresponse Rates

The nonresponse rate is defined as the percentage of all potentially eligible units (or elements) that do not have responses to—at least a certain proportion of—the items in a survey questionnaire. Thus, a nonresponse rate can be calculated at the unit level (in which all data from a sampled respondent are missing) and/or the item level (in which only data for certain variables from a sampled respondent are missing).

The nonresponse rate is not the same as nonresponse error. Nonresponse error occurs when nonrespondents in a survey are systematically different from respondents in nonnegligible ways that are germane to the objects of the study. For example, if the survey attempts to assess public opinion on the president's new plan for the Iraq War, then nonresponse error would occur if citizens who express their opinions on the issue were significantly more likely to oppose the plan than were those who were sampled but from whom no data were gathered. As such, the nonresponse rate alone does not tell whether there is nonresponse error or how much error exists in the survey. However, knowing the nonresponse rate is a necessary step toward estimating the nonresponse error.

All sampled cases in a survey can be categorized into one of four major groups: (1) eligible cases with sufficient data to be classified as *responses*, (2) eligible cases with no sufficient data (i.e., *nonresponses*), (3) cases of *unknown eligibility*, and (4) *ineligible cases*.

Responses

Cases that are treated as responses can also be divided into two groups: complete responses and partial responses. Standards that are widely used to define partial responses versus complete responses include the following: (a) the proportion of all applicable questions completed; (b) the proportion of essential questions completed; and (c) the proportion of all questions administered. Essential questions may vary, depending upon the purposes of a survey. In the case of the previously mentioned survey about public approval of the new war plan, the crucial variables may include attitudes toward the plan, party identification, and so on. Survey researchers should define what constitutes a complete versus a partial response based on one of the standards or a combination of the standards prior to data collection. Complete responses, for example, could be defined as cases that have answers to 95% or more of essential questions, whereas partial responses might be defined as cases with answers to 50–94% of such questions. Cases with fewer than 50% of the essential questions answered could be treated as *breakoffs*, which is a form of nonresponse. By this definition, only cases with data for at least 50% of crucial questions would be deemed as responses or partial responses.

Nonresponses

Nonresponses can result from noncooperation or refusals, noncontacts, and from other factors. The situations in which instances of nonresponse occur vary across surveys and across different sampling or data collection modes. Refusals happen in random-digit dialing (RDD) telephone or in-person surveys of households when a household has been contacted and either a responsible household member or the designated respondent has refused to participate in the study. Breakoffs refer to a premature termination from an interview. In mail or online surveys, refusals occur when contact has been made with a specifically named respondent or with a household or organization where the respondent works or lives, and the respondent or a responsible member of the household or organization has refused to participate. However, researchers often are not certain when this has happened as they typically receive no direct evidence to this effect. A breakoff occurs when partially completed questionnaires are returned with some notification suggesting that the respondent refused to complete the rest of the questionnaire. The distinction between breakoffs and partial responses should be pre-determined by researchers using the criteria defining a complete response versus a partial response, as previously mentioned.

Noncontacts in RDD telephone surveys of households refer to cases in which a telephone number is confirmed as a residential household, but the selected member of the household is never available to speak with an interviewer. In in-person surveys, noncontacts happen when the interviewers cannot reach the respondents, either because they cannot enter the building where the respondents live or work or because the respondents are not there and available at the time. In the case of mail or online surveys, noncontacts refer to the cases in which a questionnaire is returned after the deadline or researchers are notified that the respondent is not available during the time when the study is in the field. Here also, researchers rarely have firm evidence that this, if fact, has occurred. In mail and Internet surveys, there are relatively few instances in which the researchers learn that the sampled respondent was never contacted (e.g., receiving a notification from the postal service or an Internet server that the packet or email was undeliverable).

Other types of nonresponse include cases in which the respondent is or was eligible and did not refuse to participate in the study, but no data were collected for reasons of health, language, literacy, and so on.

Unknown Eligibility

Unknown eligibility occurs in RDD telephone surveys of households when it is not known if a sampled telephone number belongs to a residential household or whether there is an eligible respondent living in the household. In in-person household surveys, a case is categorized as unknown eligibility when it is unknown whether an eligible household is located at the sampled address or whether an eligible respondent lives in the place. In mail or online surveys, unknown eligibility includes cases in which nothing is known about whether a questionnaire has reached the respondent or whether the respondent is eligible.

Ineligible Cases

In landline RDD telephone surveys of households, ineligible cases may include phone numbers of households that are located outside of the geographical area of interest. For example, researchers may be interested in

Wisconsin residents' opinions about the seat belt law, but some telephone numbers selected by the computer may reach residents of Illinois or Minnesota and therefore should be considered as ineligible. Often with U.S. cell phone RDD samples this is even more problematic. Ineligible cases in RDD surveys also include telephone numbers of fax machines, nonresidential households, and/or out-of-service numbers. In in-person household surveys, ineligible cases result from the situation in which there is no household located at the sampled address or no eligible respondents within a sampled household. A case is deemed as ineligible in mail or online surveys when the researchers have evidence that a respondent fails to pass the screening questionnaire designed to assess eligibility.

Formulae to Calculate Nonresponse Rates

The calculation of nonresponse rates varies depending upon what are considered as potentially eligible units. The first formula of nonresponse rate (NRR1) generates the minimum nonresponse rate because the potential eligible cases include all cases of responses, all cases of nonresponses, and all cases of unknown eligibility:

$$NRR1 = nonresponses/(responses + nonresponses + unknown\ eligibility).$$

The second way of calculating the nonresponse rate (NRR2) requires researchers to estimate the proportion of cases with unknown eligibility (separately by type, e.g., for answering machines, for ring–no answers, for busy signals) that may be actually eligible—that is, the *e* term in the formula NRR2—based on the best available scientific evidence. By doing so, researchers can make a more accurate estimate of eligible units, thereby having a better estimate of the nonresponse rate. When using NRR2, the evidence used for the estimation must be provided in detail. For instance, researchers might know that 40% of cases with "ring–no answer" in an RDD telephone survey actually are eligible according to past survey experience. Then they can set the value of *e* as 0.4 for the ring–no answer final dispositions and calculate the rate using the estimate.

$$NRR2 = nonresponses/(responses + nonresponses + [e \times unknown\ eligibility]).$$

A third type of nonresponse rate calculation (NRR3) does not treat cases of unknown eligibility as potentially eligible cases. Thus, NRR3 can be considered as a special case of NRR2, in which *e* equals zero (i.e., it is assumed that there are no eligible cases among those whose eligibility is unknown). NRR3 may also be used in rare situations in which the eligibility of all cases is known. This formula generates the maximum nonresponse rate because the denominator is the smallest among the three computations:

$$NRR3 = nonresponses/(responses + nonresponses).$$

Complex surveys require more complicated ways to calculate nonresponse rates. There are three general situations of complex design, including (1) single samples with unequal probabilities of selection, (2) multi-wave panels, and (3) surveys that use a listing from a previous survey as a sample frame. In single-stage designs in which the units are sampled with unequal probabilities, the rates should be weighted by base weights that are the inverse of the selection probabilities or a number that is proportional to the inverse. In multiple-wave designs, the rates for the units that are sampled at the last stage should incorporate nonresponse at the earlier stages. For example, a three-wave, longitudinal survey should report both the nonresponse rate for the third wave and the cumulative nonresponse rate across the three waves. In two-phase designs that subsample respondents from a previously existing frame, the nonresponse rate of the current survey should be reported, and so should be the nonresponse rate that incorporates the previous sample (i.e., one that calculates nonresponse from both the current and the previous sample).

Weiyu Zhang and Xiaoxia Cao

See also Completed Interview; Complex Sample Surveys; *e*; Item Nonresponse; Noncontact; Nonresponse; Nonresponse Bias; Nonresponse Error; Partial Completion; Refusal; Response Rates; Standard Definitions; Unit Nonresponse

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Curtin, R., Presser, S., & Singer, E. (2002). Changes in telephone survey nonresponse over the past quarter-century. *Public Opinion Quarterly*, *69*, 87–98.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). *Survey nonresponse*. New York: Wiley.

Lavrakas, P. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

NONSAMPLING ERROR

Nonsampling error is a catchall phrase that refers to all types of survey error other than the error associated with sampling. This includes error that comes from problems associated with coverage, measurement, nonresponse, and data processing. Thus, nonsampling error encompasses all forms of bias and variance other than that associated with the imprecision (variance) inherent in any survey sample.

Coverage error refers primarily to the bias and variance that may result when the sampling frame (the list from which the sample is drawn) used to represent the population of interest fails to adequately "cover" the population, and the portion that is missed differs in nonignorable ways from the portion that is included on the frame. Coverage error also includes bias and variance that can result when a within-unit respondent selection technique is used that does not adequately represent the population at the level of the individual person. *Nonresponse error* refers to the bias and variance that may result when not all those who are sampled have data gathered from them, and these nonresponders differ in nonignorable ways from responders on variables of interest. Item-level nonresponse error includes the bias and variance that may result from cooperating respondents who do not provide answers (data) to all the variables being measured if the data they would have provided differ in nonignorable ways from the data that the other respondents are providing on those variables. *Measurement error* refers to bias and variance that may result related to the questionnaire, the behavior of the person who gathers the data, the behavior of respondents, and/or the mode of data collection. *Data processing errors* refer to the bias and variance that may result from mistakes made while processing data, including the coding and recoding of data, the transformation of data into new variables, the imputation of missing data, the weighting of the data, and the analyses that are performed with data.

Researchers concerned with nonsampling error can take two different strategies to try to deal with it. First, they can implement numerous methodological and other quality control techniques to try to reduce the amount of nonsampling error that results in their studies; this typically adds to the costs of the research study. Second, they can build in methodological studies to try to measure the nature and size of the nonsampling errors that cannot be reduced to negligible levels; this also adds to the project cost but often not as much as the first approach.

Many people appear to think that nonsampling error applies only to research studies that use the survey method of data collection. However, each type of error that makes up nonsampling error has its counterpart in any form of social research, be it qualitative or quantitative, including experiments and quasi-experiments, content analysis, observational research, cognitive interviewing, and focus groups.

Paul J. Lavrakas

See also Bias; Content Analysis; Coverage Error; Measurement Error; Nonresponse Error; Sampling Error; Sampling Frame; Total Survey Error (TSE); Variance

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

NONTELEPHONE HOUSEHOLD

Telephone surveys became an acceptable mode of data collection in the United States in the 1970s, when approximately 90% of households in the United States had a telephone. According to the 2000 Census, 98% of U.S. households contained a telephone. However, the 2000 Census did not distinguish between wireline and wireless service, so having a "telephone" could mean having a landline phone, a wireless phone, or both. In each year since the 2000 Census, more and more households began to substitute wireless telephone service for their wireline or landline telephone service. This phenomenon, often referred to as "cutting the cord," has introduced additional coverage bias in traditional wireline random-digit dial (RDD) samples, as in 2008 approximately 20% of U.S. households had only a cell phone.

By 1986, only a little more than 7% of households were without a telephone. Analysis of National Health Interview Survey (NHIS) data by Owen Thornberry and James Massey in 1988 showed that certain sociodemographic cohorts were disproportionately represented among the nontelephone population, particularly families and persons living in poverty. If information on low income, or any correlates of low income, was an important objective of a telephone survey, researchers were encouraged not to use a telephone sample for data collection. As the number of households without a telephone declined into the 1990s, concerns about bias diminished but did not disappear.

In 1995, Scott Keeter proposed a method for minimizing the bias associated with the exclusion of households without a telephone. Keeter showed that interruption of telephone service is usually episodic in nature. Based on this observation, he proposed using survey data collected from respondents reporting interruption of telephone service to make adjustments for the nontelephone population.

Based on the 2006 NHIS data and data from Mediamark's 2006 national syndicated survey, only 2% of households have no phone of any kind. Findings from the 2007 NHIS survey show 15% cell phone only and another 10% as being "cell phone mostly" (i.e., they have a landline and a cell phone but essentially do not use their landline for incoming calls). This shift in telephone ownership away from landline telephones means that a traditional RDD sample today will represent only 87% of all households. In 2006, E. Deborah Jay and Mark DeCamillo experimented with a Keeter-like approach for adjusting for cell phone only households. In their study, they asked about interruption of telephone service and whether the household had cell phone service during that interruption. As with the Keeter study, the demographics of recent cell phone only households were similar to published demographics of cell phone only households. The authors proposed using data collected from respondents that reported a recent cell phone only status to adjust estimates for noncoverage of the cell phone only population.

Including households with no telephone service of any kind is not a viable option for RDD surveys. However, cell phone only households can be included in telephone surveys. Frames exist for sampling cell phone numbers, and there has been a significant amount of ongoing research related to sampling cell phones both in the United States and around the world. Many methodological issues remain to be solved, and in the United States sampling cell phones is uniquely complicated by Telephone Consumer Protection Act (TCPA) regulations that prohibit dialing a cell phone with an autodialer.

Linda Piekarski

See also Cell Phone Only Household; Cell Phone Sampling; Dual-Frame Sampling; Telephone Consumer Protection Act of 1991; Telephone Households; Telephone Penetration; Telephone Surveys

Further Readings

Frankel, M. R., Srinath, K. P., Hoaglin, D. C., Battaglia, M. P., Smith, P. J., Wright, R. A., et al. (2003). Adjustments for non-telephone bias in random-digit-dialing surveys. *Statistics in Medicine*, *22*, 1611–1626.

Jay, E. D., & DeCamillo, M. (2006, May). *Identifying recent cell phone-only households*. Paper presented at 2006 Annual Conference of the American Association for Public Opinion Research, Montreal, Canada.

Keeter, S. (1995). Estimating telephone non-coverage bias with a telephone survey. *Public Opinion Quarterly*, *59*(2), 196–217.

Kim, S. W., & Lepkowski, J. L. (2002). Telephone household non-coverage and mobile telephones. *2002 Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 1845–1850). Alexandria, VA: American Statistical Association.

Lavrakas, P. J., Steeh, C., Boyle, J., Blumberg, S., Callegaro, M., et al. (2008). *Guidelines and considerations for survey researchers when planning and conducting RDD and other telephone surveys in the U.S. with respondents reached via cell phone numbers*. Lenexa, KS: American Association for Public Opinion Research. Retrieved June 1, 2008, from http://www.aapor.org/uploads/Final_AAPOR_Cell_Phone_TF_report_041208.pdf

Thornberry, O. T., & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 25–49). New York: Wiley.

Nonverbal Behavior

Nonverbal behavior is physical action that complements, supplements, or takes the place of spoken words or sounds. Examples include, but are not limited to, facial expressions, body postures, and gestures.

Data collection in survey research may include the cataloging (observing and coding) of nonverbal behavior in order to help put verbal data—written or spoken words or other utterances—into context for more in-depth analysis and interpretation. For example, a study based on personal interviews of people who are reluctant to speak with the interviewer may gauge the level of rapport that the interviewer was able to establish by examining the nonverbal behavior of respondents. If these respondents are frowning, turning their backs to the interviewer, or otherwise demonstrating discomfort, this information could be used to gauge the credibility of the answers of those interviewed.

Nonverbal behavior can change, alter, enhance, supplement, complement, or contradict the meaning of an act of verbal communication. Nonverbal behavior sometimes contradicts or confuses the meaning of verbal communication. For example, if an individual tells an interviewer that they enjoy an activity but they frown as they say so—the frown communicates something different than a smile or a neutral expression. In this case, the respondent may be providing a socially desirable yet inaccurate verbal answer. A researcher could work toward producing more valid data by recording the nonverbal behavior that accompanies the verbal communication.

Survey researchers also need to be concerned about the nonverbal behavior that interviewers exhibit. Interviewers who are expected to remain neutral while administering a questionnaire may keep their voice and language neutral but may inadvertently demonstrate nonverbal signals that could bias a respondent's answers. Thus, interviewer training, especially in the case of in-person (face-to-face) interviews, should emphasize interviewers being aware of their own nonverbal behavior.

Nonverbal communication and its meanings can vary across cultures and groups. For example, whether a speaker looks at a listener in his or her eyes during conversation has different meanings in different cultures. In some cultures, looking into the eyes of the person to whom you are speaking is considered respectful and desirable behavior. In other cultures, a direct gaze is thought to be a sign of disrespect. In addition, recent studies of eye gazing and human communication have found that eye gazing may not be only a cultural phenomenon but also a physiological one. People with differing brain structures—for example, some people with autism spectrum disorders—look at areas on faces other than the eyes because direct eye gaze can provoke a physiological fear response for them.

The conditional and contingent meanings of nonverbal behavior as a complement to verbal communication can make coding and analyzing the information a challenging, but potentially very useful, aspect of research.

Heather H. Boyd

See also Attitude Measurement; Behavior Coding; Coding; Face-to-Face Interviewing; Interviewer Neutrality; Interviewer Training; Measurement Error; Social Desirability

Further Readings

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to unmasking the emotions from facial cues.* Cambridge, MA: Malor Books.

Null Hypothesis

A null hypothesis is one in which no difference (or no effect) between two or more variables is anticipated by the researchers. This follows from the tenets of science in which empirical evidence must be found to disprove the null hypothesis before one can claim support for an alternative hypothesis that states there is in fact some reliable difference (or effect) in whatever is being studied. The null hypothesis is typically stated in words to the effect that "A equals B." The concept of the null hypothesis is a central part of formal hypothesis testing.

An example in survey research would be a split-half experiment that is used to test whether the order of two question sequences within a questionnaire affects the answers given to the items in one of the sequences, for example, in crime surveys where fear of crime and criminal victimization experience are both measured. In this example, a researcher could hypothesize that different levels of fear would be reported if the fear items followed the victimization items, compared to if they preceded the victimization items. Half the respondents would be randomly assigned to receive one order (fear items, then victimization items), and the other half would receive the other order (victimization items, then fear items). The null hypothesis would be that the order of these question sequences makes no difference in the answers given to the fear of crime items. Thus, if the null hypothesis is true, the researcher would not expect to observe any reliable (i.e., statistically significant) difference in levels of fear reported under the two question ordering conditions. If results indicate a statistically reliable difference in fear under the two conditions, then the null hypothesis is rejected and support is accorded to the alternative hypothesis, that is, that fear of crime, as reported in a survey, is affected by whether it precedes or follows victimization questions.

Another way of understanding the null hypothesis in survey research is to think about the crime survey example and the confidence intervals that can be calculated around the fear of crime measures in the two ordering conditions. The null hypothesis would be that the 95% confidence intervals for the fear measures under the two orders (conditions) would overlap and thus not be reliably (significantly) different from each other at the .05 (alpha) level. The alternative hypothesis would be that the confidence intervals would not overlap, and thus the fear measures gathered under one order are reliably different from the same fear measures gathered under the other order.

Rejecting a null hypothesis when it is in fact true is termed a *Type I error*. Not rejecting a null hypothesis when it is fact false is termed a *Type II error*.

Paul J. Lavrakas

See also Alpha, Significance Level of Test; Alternative Hypothesis; Confidence Interval; Experimental Design; *p*-Value; Split-Half; Statistical Power; Type I Error; Type II Error

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth/Cengage Learning.

Campbell, D. T., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Number Changed

When a telephone survey is conducted, some dialings will result in a message stating that the number they have dialed has been changed. The message often includes the new number. "Number changed" dispositions in RDD samples are normally classified as ineligible, but there are some circumstances for which the researcher might want to considered such numbers eligible and call the new number.

When a household changes its landline phone number, it is usually because that household has moved to a new location. If the move was to a location outside the exchange boundaries of the old telephone number, a U.S. household traditionally had to obtain a new telephone number in an exchange serving their new address. However, number portability has changed this by allowing people to keep their telephone number when they move. A household also might change its number without relocating. For example, they might want to replace their directory-listed number with a new, unlisted number. Or a household might elect to change their service provider to a competitive local exchange carrier or to a provider of Voice over Internet Protocol (VoIP) services. In areas where number portability is not available, the household would be required to get a new telephone number in order to change service providers.

Usually numbers that result in a "Number changed" message are removed from a random-digit dialed (RDD) sample because the new geographic location makes them ineligible for the survey. A more important reason for their exclusion is that an RDD sample could have allowed both numbers—the old number and the new number—to be eligible for selection. If researchers elect to call the new number, they must also apply a weight to correct for the multiple probabilities of selection of that household.

However, for telephone surveys that sample specifically named persons—for example, when working from a client-supplied list of customers—it often will be appropriate to call the new number. Other instances in which it would be appropriate to call a changed number would be (a) recontacting a respondent for a follow-up interview and (b) conducting longitudinal surveys of the same respondents. In such cases, multiple probabilities of selection usually would not be a concern.

Linda Piekarski

See also Dispositions; Final Dispositions; Hit Rate; Number Portability; Random-Digit Dialing (RDD); Telephone Surveys

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Retrieved April 21, 2008, from http://www.aapor.org/uploads/standarddefs_4.pdf

Number Portability

Number portability is the ability of users of telecommunications services in the United States and most other countries to keep their existing telephone number when changing from one local service provider to another. Under Federal Communications Commission (FCC) rules, the implementation of local number portability (LNP) in the United States has caused some problems for telephone survey researchers. In the early days of wireline-to-wireline portability, both numbers associated with a porting request could be dialed successfully. When wireline-to-wireless porting was implemented, there were concerns that this would influence the ability of researchers and telemarketers to comply with the portion of the Telephone Consumer Protection Act (TCPA) of 1991 that limited calls to certain types of phone numbers, including wireless phone numbers.

In 1996, the U.S. Congress amended the Telecommunications Act of 1934 to establish a framework that would promote competition and reduce regulation in all telecommunications areas. It was recognized that certain barriers to competition would need to be eliminated, specifically the inability of customers to switch from one service provider to another and retain the same phone number. New FCC rules were enacted that gave consumers the ability to switch from one service provider to another and keep their existing telephone numbers. The rules were applicable only locally, that is, only within a local exchange or rate center. If a person moved from one geographic area to another, the number would not be portable. Telephone companies were allowed to charge a fee to cover their porting costs. However, because of these costs, most small landline companies were not required to port numbers to wireless carriers until the FCC had completed a study about the effects of porting rules on small companies.

Local number portability in the United States was implemented in phases. Portability between local landline (wire) service providers was implemented in 1998. The FCC had also required that three categories of CMRS (commercial mobile radio service) providers—cellular providers, broadband personal communications service (PCS) providers, and covered specialized mobile radio (SMR) providers—also provide number portability. The commission concluded that requiring them to do so would promote competition between and among local landline and wireless services. A separate timetable for compliance was established for CMRS providers. LNP between wireless service providers was finally implemented in November 2003 and between landline and wireless services in early 2005.

Porting in the United States requires two 10-digit numbers. One is the original subscriber number, which is no longer a valid local routing number or switch, and the other is a number in a prefix or 1000-block belonging to the new carrier that is used for rerouting a call to the correct end-user location and for accounting. A 1000-block is a block of 1,000 consecutive local numbers within a prefix, all starting with the same seven digits (e.g., 203-255-1XXX). Within a prefix, 1000-blocks can belong to different service providers offering different types of service. Since the subscriber's new carrier can also provide telephone service to new or existing customers as well as customers that are changing carriers, these remote porting numbers can and do occur in 100-blocks found on list-assisted RDD databases. A 100-block is a block of 100 consecutive numbers starting with the same eight digits (203-333-65XX). List-assisted RDD frames normally contain only 100-blocks with one or more directory-listed numbers from which RDD samples are generated by appending a random number between 00 and 99 to selected 100-blocks. Once a new subscriber (e.g., 203-333-6555) appears in a telephone directory, the 100-block 203-333-65XX becomes eligible for RDD samples. During the generation of an RDD sample, both kinds of numbers—those belonging to regular subscribers and those belonging to subscribers that have ported their number to this new carrier—can be generated during random number assignment.

As part of local number portability in the United States, a special database, currently maintained by NeuStar (the FCC-designated administrator of the U.S. database of ported numbers), and software were developed for telecommunications carriers to use. The database contains each ported number and the new number associated with that ported number. The software allows any telecommunications carrier to query the database before connecting a telephone call. If the number a customer has dialed is on the database, the carrier will be provided with the information required to redirect that call to the correct switch and location routing number (LRN). In the early years of LNP, the required databases and software were not in place everywhere. During that period, portability was sometimes handled using traditional call-forwarding techniques. Since both numbers, the number being ported and the number associated with the new switch, could be randomly generated in an RDD sample, both numbers would be "working" and would connect to the same residence. This caused what has been frequently referred to as *ghost numbers*, the telephone numbers associated with the new service provider but not recognized by a respondent to a survey.

Ghost numbers created two problems for researchers using RDD samples. First, a household that had ported its number would only report one telephone number, the number that they ported to a different provider. But in reality they had two numbers, the number they had ported and the number belonging to their new service provider. This meant that they had two chances of selection. Second, if contact was made on the telephone number belonging to the new service provider, the respondent would not recognize that number as being their telephone number. This meant that it was unclear to an interviewer whether the number had been dialed incorrectly or whether the number was a ghost number. Although call-forwarding techniques may still be used in areas that are transitioning to local number portability, the vast majority of competitive LRNs will not connect if dialed directly. However, their presence as disconnects or nonworking numbers in RDD samples will contribute to reduced sample productivity.

By 2005, full portability had been implemented in the United States, allowing porting between wireless and landline services. At this point, telephone survey research was further affected because a landline telephone number ported to a wireless phone number might easily appear in an RDD landline sample. This was a particular concern, because dialing wireless numbers in the United States using automated telephone equipment is a violation of TCPA regulations. The FCC was lobbied to provide a method for identifying these numbers. NeuStar, and the FCC agreed to license this information for an annual fee. Licensees have access to two files of ported numbers that are updated daily: a file of wireless-to-wireline telephone numbers and a file of wireline-to-wireless.

According to data from the FCC, by December 2004, 30.5 million numbers had been ported from one landline carrier to another, and more than 10 million numbers had been ported from one wireless carrier to another. By contrast, only 818,000 numbers had been ported from landline service to wireless service and only 10,000 from wireless service to landline service. Based on the NeuStar databases, as of July 2006, 2.4 million landline numbers have been ported to a wireless service provider, while only about 90,000 wireless numbers have been ported to landline service.

The porting of numbers between landline and wireless service providers can create coverage bias in RDD frames. In order to protect clients from inadvertent violation of the TCPA regulations, most sample providers routinely remove numbers ported to a wireless service and known wireless numbers (which would include those that have ported their wireless number to landline service) from their databases and telephone samples. Conceptually, the wireless-to-wireline file and the wireline-to-wireless file could be used to eliminate this bias. However there are strict prohibitions on the use of these files. Subscribers are contractually obligated to use this information solely to "meet the requirements and conditions of the TCPA." Therefore, using these numbers to augment a telephone frame could be in violation of these restrictions, unless they are pre-identified and hand-dialed by a qualified subscriber.

Ongoing analysis of this file by Survey Sampling International suggests that the coverage bias is minimal. On average, only about 14% of these ported numbers appear in a list-assisted frame, representing only 0.03% of the total frame. This file changes on a daily basis. On the other hand, research firms or call centers usually dial telephone sample numbers for days or weeks following their selection; researchers are encouraged to subscribe to the service and do their own scrubbing on a daily basis if they want to avoid all legal issues.

Another serious problem caused by number portability that U.S. telephone surveys face is the fact that some numbers for area codes and prefixes in a particular local area may reach people living in other geographic areas. This can happen when a landline number has been ported to a wireless service provider or a provider of Voice over Internet Protocol (VoIP). In these cases the cell phone or VoIP router and/or computer can be *transported* from one geographic location to another. Additionally, some VoIP service providers offer subscribers a telephone number (for inbound calls only) in a totally different area code. As the frequency of this transporting increases, telephone surveys will need to devise cost-effective ways to screen for the geographic eligibility of the household or respondent.

Linda Piekarski

See also Access Lines; Call Forwarding; Cell Phone Only Household; Cell Phone Sampling; Geographic Screening; Prefix; Random-Digit Dialing (RDD); Telephone Consumer Protection Act of 1991

NUMBER VERIFICATION

The verification of a telephone number in a telephone survey is done by an interviewer who confirms with a respondent that the number that was ostensibly dialed is in fact the number that was reached. The need to do this has been reduced over the years as more landline telephone surveys have come to be conducted with computer-assisted telephone interviewing (CATI) systems that include software and hardware to place the calls to the sampled telephone numbers, as opposed to having interviewers manually dial the numbers. However, the need for number verification has not been eliminated completely, as even with automatic dialing equipment mistakes sometimes occur. Furthermore, in the United States and due to current federal telecommunications regulations, all cell phone numbers that are sampled for a telephone survey must be hand-dialed—unless the cell phone owner has given the survey organization prior consent to be called—and thus interviewers will need to verify whether they have dialed the sampled number correctly.

There are several reasons that a landline (or cell phone) telephone survey may not reach the correct number, even when using equipment to place the calls. For example, national and local telephonic systems are subject to occasional error when "wires get crossed" (the electronic signals get mixed up), thus leading to a call reaching another number than the one to which it was intended. Call forwarding, whereby one telephone number is programmed by its owner to ring at another number, can also lead to the "wrong" number being reached. This is problematic when a business number that appears in an RDD sample is forwarded to a residential number. In this case, the household would not be eligible unless one of its residential numbers also was sampled. However, if a business number were reached because the residential number that was sampled in RDD was forwarded to it, then the resident reached via her or his forwarded home number would remain eligible. If the telephone survey has sampled "named persons," then reaching them on a different number than what was dialed does not make them ineligible. In these instances of number verification, the interviewer will learn whether the number dialed is not the number reached, and if it is, then the interview may continue or may be politely terminated and coded as out-of-sample depending on the manner of forwarding that took place. Using manual dialing will lead to errors made by the interviewers who are placing the calls. Thus, whenever interviewers are hand-dialing sampled numbers, a verification that the correct number was reached always should be included in the introduction of the survey before the questionnaire is administered. Typically, this verification is done after cooperation has been gained, because it is assumed that doing it too soon after initial contact will lead to an increase in nonresponse.

Paul J. Lavrakas

See also Call Forwarding; Cell Phone Sampling; Out of Sample; Random-Digit Dialing (RDD); Standard Definitions; Telephone Surveys

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

O

Open-Ended Question

The selection of question structure is fundamental to the process of questionnaire construction. The open-ended question is one type of structure; the other, more commonly used alternative is the closed-ended question. The open-ended question does not provide answer categories. The person (respondent) who is asked an open-ended question formulates the answer and gives the response in his or her own words. Although this structure gives the respondent more freedom in crafting an answer, it also increases the cognitive effort. Without answer choices as cues to aid in understanding the question and deciding on an answer, the respondent has to perform additional cognitive tasks before he or she responds.

Reasons to Use Open-Ended Questions

All open-ended questions are alike in that the respondent is not given answer choices. However, the reasons for using this structure and the level of cognitive effort needed to respond can vary. The following are seven examples that illustrate different reasons for open-ended questions.

1. *Build rapport and encourage participation*. Asking an easy-to-answer question at the beginning of the questionnaire signals expressing an opinion as a benefit of survey participation and acknowledges the importance of what the respondent has to say. Sometimes initial qusestions used for this purpose are considered "warm-up" questions because one of the main objectives is to engage the respondent (e.g., *In your opinion, what is the most important issue facing the United States today?*).

2. *Get factual information*. When there is a wide range of answers expected to provide individual factual information, an open-ended structure can address the problem of having a list of more response choices than it is practical to include in a questionnaire. Factual information may be a request for a verbatim or for a numeric response (e.g., *What is your occupation? How much do you plan to spend on holiday gifts? Last year, what was your total annual household income?*).

3. *Expand a list*. When a closed-ended question offers a list of specific items or response choices (e.g., places where people get their news), a follow-up question asking about additional information can ensure that the pre-listed choices have not omitted any options (e.g., *Are there any others?*).

4. *Explain a prior answer*. An open-ended question can deepen the understanding of the response to a preceding question by obtaining additional details on the reason for the answer choice (e.g., *Why were you satisfied/dissatisfied with your last doctor's appointment?*).

5. *Establish knowledge*. A test question can distinguish between more and less informed respondents to enhance the understanding of opinion formation (e.g., *Who are the U.S. senators from your state?*).

6. *Clarify terminology*. Asking respondents to define a key word in a question documents their level of understanding. It can also inform the variation in the meanings of words used among the respondents

who gave an answer (e.g., *What does* welfare *mean to you?*).

7. *Explore new topics*. The questionnaire can be an opportunity to get suggestions for future survey topics that are especially salient to the respondent. In particular for longitudinal studies, this information can inform the development of future questionnaires (e.g., *Questionnaires by their nature are limited. What city services, not included in this questionnaire, should be evaluated in future surveys?*).

Data Quality Considerations

When an open-ended question is used, particular attention must be paid to other aspects of the survey process that can affect the data quality and are specifically related to this structure: method of data collection, coding verbatim responses, and time and expenditure trade-offs.

Method of Data Collection

There are two basic methods used to collect information: self-administered (the respondent self-records answers by writing on a paper questionnaire or by entering a response into a computer) and interview (the respondent answers questions that are read and then recorded verbatim by another person, i.e., the interviewer). With the self-administered method, the respondent is responsible for providing a quality answer. The types of respondent-related errors specific to the open-ended structure are (a) missing answers; (b) incomplete responses; (c) misunderstood terminology; and (d) illegible writing. The format of a self-administered paper or electronic questionnaire can provide some assistance in reducing these errors. For example, the size of the space provided is a visual cue on how much information the respondent is expected to report—a smaller space results in less information, while more information is provided when there is a larger space. When there is a request to write in numeric factual information (e.g., *What was the last date you saw your doctor?*), clear instructions on how to provide the month, day, and year will reduce the variation, and possible errors, on how respondents report this information.

An interview can reduce some of the self-administered errors because an interviewer guides the respondent; however, there are other data quality considerations that need attention. The types of errors that can result from interviewer effects are biased answers as result of leading probes; answers given in order to provide socially desirable responses; and inaccurate verbatim recording. Specific interviewer training on how to ask open-ended questions is essential to minimize these types of errors.

In general, compared to the self-administered mode, more complete and accurate information can be expected when an interview method is used. When a survey includes both methods to give respondents a choice on how to participate, the variation in responses associated with each method needs to be considered.

Coding Verbatim Responses

Coding verbatim responses is necessary with open-ended questions. While one of the main advantages of using an open-ended structure is getting specific, individual information, the lists of verbatim answers need to be organized to be useful for data analysis and reports. Developing numeric codes to accurately represent the verbatim responses is challenging. The quality of open-ended data is diminished when careful attention is not given to code development. Errors can also occur when a person reads a verbatim answer and has to make a judgment about the most appropriate code to assign. Thorough training on how to make these judgments improves the accuracy and the confidence in reliable coding results. When multiple people are coding verbatim responses, the quality of the data also depends on intercoder reliability. To minimize the amount of coding that is needed on a questionnaire completed with an interviewer, a list of precoded answers can be provided. While the question is still asked using an open-ended structure, the interviewer uses the precoded list to classify a verbatim answer. There are two possible sources of error associated with a list of precoded choices: the reliability of the interviewer's judgment in selecting the appropriate answer and the accuracy of the items on the precoded list.

To obtain quality information, open-ended questions require sufficient time and financial resources to support the actions required for data quality. Time needs to be allowed when the questions are being answered and for coding the answers after the completed questionnaires are returned. A typical open-ended question can take two or three times longer for a respondent to complete than a closed-ended question because of the cognitive process associated with formulating a response and the extra time required to record a complete, legible verbatim answer. Also, because coding is needed to organize and quantify open-ended questions for data analysis and reporting, additional time has to be allocated for this process. Including open-ended questions means additional funds are needed to provide training for interviewers and coders and for professional time used to develop a coding system and to monitor the quality of the codes.

Janice Ballou

See also Closed-Ended Question; Coding; Interviewer Effects; Interviewer-Related Error; Mode Effects; Precoded Question; Questionnaire Design; Questionnaire Length; Respondent Burden; Respondent-Related Error; Verbatim

Further Readings

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.
Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.
Dillman, D. A. (2000). *Mail and Internet surveys*. New York: Wiley.
Payne, S. P. (1951). *The art of asking questions*. Princeton, NJ: Princeton University Press.
Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Opinion Norms

From a social science perspective, public opinion is much more than an aggregation of polling statistics. While individual survey responses are an instrumental component of understanding public opinion as a social force, the context in which the individual operates (e.g., media environment, typical discussion patterns) is an equally important consideration in obtaining a better understanding of the evolution of public opinion. Recognizing the normative aspects of a public opinion climate allows researchers to understand better how individuals come to possess opinions and how those opinions are shared with others.

Past work on behavioral norms offers insight as to how contextual forces, or "climates of opinion," can influence the actions and expressions of group members. Such social norms can be classified into two main categories, descriptive and injunctive. *Descriptive* norms are informational and describe the way things *are* within a given social setting, whereas *injunctive* norms possess a sanctioning function and prescribe the way things *ought* to be. Individuals who violate injunctive norms (i.e., engage in *proscribed* behavior) run the risk of alienating themselves from those around them.

By making use of interactions with important reference groups and exposure to available media, individuals are able to get a sense of what is socially acceptable when expressing political views and opinions. This iterative process establishes the normative environment surrounding opinion expression—it is within this climate that individuals may feel more or less inclined to express their own view. Respondents' perceptions of congruity between their own opinion and the perceived opinion of a given reference group can either encourage or dissuade opinion expression, much like behavioral norms influence the actions of those within a given social context.

Individual perceptions of this climate are an important and quantifiable aspect of understanding the impact that normative environments can have on individual behavior and expression. Aggregated perceptions of public opinion, such as those generated from survey data, constitute a "social barometer" that quantifies both the extremity and the amount of agreement among actors as related to these prevailing social forces.

Researchers interested in capturing the normative aspects of public opinion should account for the following opinion characteristics: (a) the valence and the strength of individual opinion (e.g., *Do you approve or disapprove of* X?, and *To what degree?*); and (b) perceptions of the valence and strength of the group opinion (e.g., key reference groups, such as members of your community, residents of this state). Within a survey context, questionnaires need to be geared toward respondents' perceptions of the climate of opinion and can included questions such as, *In your judgment, what would the reaction be if someone expressed* strong support *for* Candidate X *during the course of a conversation among people in your neighborhood: Very positive, somewhat positive, neutral, somewhat negative, or very negative?* While survey items such as these tap into individual perceptions of the context in which opinions are expressed, they also allow for simple (experimental) manipulations of key variables within the question (e.g., substituting candidates or discussion topics, level of support or opposition, and reference groups named within the question).

With this response data in hand, Jay Jackson's Return Potential Model (RPM) provides one example for quantifying the normative opinion climate. Normative "intensity" can be evaluated using the RPM by calculating the mean group deviation from a midpoint or neutral response option. This measure specifies the approval or disapproval associated with an opinion as well as the extremity or strength with which a norm is held (e.g., slight approval versus strong approval of expressing a specified opinion). It is quantified by using the following equation:

$$I_i = x_i - m,$$

where intensity (I) is the mean deviation from the researcher's specified midpoint value (m). It is important to note that intensity is a bidirectional concept that can apply to strong disapproval as well as strong approval. In other words, the degree to which a certain opinion norm is characterized by negative intensity (i.e., an opinion is strongly opposed) is related to the probable level of social sanctions for expressing that opinion. For high-intensity norms, violation will most likely result in some form of social isolation. On the other hand, expressing opinions associated with low-intensity norms may be seen as odd but will likely bear little social cost.

When measuring the normative opinion climate, it also is important to know how much agreement exists within a specified group. The RPM model also captures normative "crystallization," which quantifies the amount of consensus associated with a norm. Specifically, this assessment of crystallization measures the level of agreement regarding what opinions are appropriate to express. Mathematically, crystallization is *inversely* associated with variance among respondents' normative views and therefore becomes greater when there is relatively more agreement about public opinion. To properly derive crystallization (C), the following equation is used:

$$C_i = 1/(x_i - x_{mean}),$$

where C_i is the inverse of the deviation from the mean approval rating among all respondents. Highly crystallized opinion norms are well understood and solidified. Low crystallization is likely to be associated with normative ambiguity. For example, if it is perceived that everyone in a certain group strongly supports Candidate X, it is safer for group members to speak highly of that candidate without fears of social sanctions. Because such a normative environment is unambiguous (i.e., highly crystallized), group members are better able to anticipate the reactions of others.

With intensity and crystallization available as quantifications of normative opinion climates, researchers are able to generate predictions related to subsequent opinion expression, willingness to deliberate, and any number of communication behaviors that may fall under the sway of public opinion as a social force.

Carroll J. Glynn and Michael Huge

See also Experimental Design; Opinions; Public Opinion; Social Isolation; Spiral of Silence

Further Readings

Glynn, C. J. (1997). Public opinion as a normative opinion process. *Communication Yearbook*, *20*, 157–183.

Jackson, J. M. (1975). Normative power and conflict potential. *Sociological Methods and Research*, *4*, 237–263.

Price, V., Nir, L., & Capella, J. N. (2006). Normative and informational influences in online political discussions. *Communication Theory*, *16*, 47–74.

Rimal, R. N., & Real, K. (2005). How behaviors are influenced by perceived norms. *Communication Research*, *32*(3), 389–414.

OPINION QUESTION

Opinions that individuals hold about an issue are potentially quite complex, considering how opinions comprise beliefs, feelings, and values on that issue. As a result, survey questions designed to assess a respondent's opinion on an issue can tap a combination of the respondent's feelings and thoughts. As composite measures, however, opinions are gauged through a variety of opinion items.

The most basic opinion question, sometimes called an "attitude question," is designed to measure the direction of opinion. That is, where does the respondent stand on the issue or attitude object? Such opinion items, typically closed-ended, can be dichotomous in nature (e.g., *Do you support or oppose abortion?* Answer categories: Support; Oppose. *Are you in favor of the death penalty for a person convicted of murder?* Answer categories: Yes; No). Opinion items also can employ a scale that measures not only the direction of one's opinion but also the respondent's intensity of opinion (e.g., *To what extent do you support or oppose abortion?* Answer categories: Strongly oppose; Somewhat oppose; Somewhat support; Strongly support).

Open-ended questions can be used as well to measure opinions, but these can be more effective in shedding light on why respondents hold the opinions they hold. For instance, respondents in the American National Election Studies who express their intention to vote for a particular candidate are asked, *Is there anything in particular about [Politician X] that might make you want to vote for him?* Other open-ended items, such as the "Most Important Problem" question (*What do you believe is the most important problem facing the nation today?*), do not offer respondents an opportunity to articulate reasons why they volunteered a particular issue. However, like the aforementioned opinion questions, responses to Most Important Problem items can be used to shape public policy, campaign efforts, or marketing strategies.

Regardless of the issue under study, how an opinion question is asked can have a significant impact on its response. Hence survey researchers routinely take into consideration a number of concerns when crafting their questionnaires. First, are nonattitudes a potential problem? Also known as "false positives" (i.e., errors of commission), nonattitudes occur when respondents report an opinion on an issue about which they know nothing or really have no attitude. To circumvent the problem of nonattitudes, survey researchers can employ a filter question immediately before the opinion question such that only respondents who report knowledge or awareness of the issue are asked their opinion about that issue. Another way to reduce nonattitudes is to offer respondents a "Don't Know" option or a middle response category (e.g., "Neither support nor oppose"). Though common, the inclusion of a middle response category can have the unintended consequence of generating "false negatives" (i.e., errors of omission), the reporting of no opinion when in fact one exists.

A second consideration in the crafting of opinion items concerns how the question and response categories are worded. The question should include a clearly specified attitude object and should not be double-barreled; put another way, respondents should not be asked to express their opinion about two attitude objects (e.g., *Do you support abortion and health care reform?*). In addition, questions should not include double negatives, colloquialisms, or leading terms such that respondents feel pressured to provide a socially desirable answer. Response alternatives, the construction of which should follow these same guidelines, also should be balanced and include a sufficient range of variation.

Content of the specific opinion question aside, survey researchers need to be mindful of the order in which the item appears in the instrument. After all, responses to opinion questions concerning a general attitude object are influenced by top-of-mind considerations. For instance, respondents asked several questions about the environment will weigh this issue more heavily than other issues when asked subsequently about how they view the president's current performance. Respondents also can try to answer opinion questions such that they appear consistent. The order of the response alternatives also can shape how individuals reply to a given question. This is important because some respondents are predisposed to select the first (*primacy effect*) or last (*recency effect*) response option. Survey researchers thus have begun to rely on rotating or randomizing response order within a given sample.

Because individuals' opinions reflect their psychological states and are measured by self-reports, the validity of responses to opinion questions is of utmost importance. Fortunately, researchers in a number of disciplines are working on theoretical and methodological fronts to better understand how responses to opinion questions are shaped.

Patricia Moy

See also Attitude Measurement; Attitudes; Attitude Strength; Closed-Ended Question; Contingency Question; Double-Barreled Question; Double Negative; Error of Commission; Error of Omission; National Election Studies (NES); Nonattitude; Open-Ended Question; Opinion Norms; Opinions; Primacy Effect; Questionnaire Order Effects; Random Order; Recency Effect; Respondent-Related Error; Response Alternatives; Response Order Effects; Self-Reported Measure; Social Desirability

Further Readings

Bishop, G. F. (2005). *The illusion of public opinion: Fact and artifact in American public opinion polls*. Lanham, MD: Rowman & Littlefield.

Glynn, C. J., Herbst, S., O'Keefe, G. J., & Shapiro, R. Y. (1999). *Public opinion*. Boulder, CO: Westview.
Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.
Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Opinions

Opinions in survey research can be defined as subjective attitudes, beliefs, or judgments that reflect matters of personal (subjective) preference. Some opinions may not be confirmable or deniable by factual evidence (e.g., a person's attitude toward the use of capital punishment), whereas others may be (e.g., the belief that a particular presidential candidate will be elected). Moreover, the strength of one's opinions may depend on one's level of knowledge or attentiveness on a subject.

The term *opinion* is often used interchangeably with *attitude* and *belief*, but opinions are a broader category that includes both attitudes and beliefs. One's subjective position on the truth of a subject is a belief, such as whether global warming is the result of human-made activity or if abstinence education lowers the rate of teenage pregnancies. Some beliefs could, at some point, be resolved with finality through science. Attitudes are a type of opinion that have an evaluative component—that is, they refer to a positive or negative evaluation of a person, idea, or object. Attitudes are latent, subjective constructs that cannot be observed directly and cannot be confirmed or denied with factual information.

Opinions are measured in surveys, which is just one way that opinions can be expressed. Opinions can also be expressed via other behaviors such as voting, participation in marches or demonstrations, or attempting to influence another person's opinion. Although surveys involve measures of individuals' opinions, surveys typically are designed to measure public opinion, which can be defined as the aggregate of opinions across the public.

Opinion Formation and Change

Opinions can be formed and changed through a number of routes including direct experience, parental influence, group determinants, elite discourse, and information from the mass media or other sources. Direct experiences strongly influence opinions and opinions based on direct experiences may be stronger than those formed via other mechanisms. Salient incidents, especially traumatic ones, can lead to indelible views on a subject, such as when a child is bitten by a dog and subsequently dislikes dogs throughout life, or when a person undergoes a religious conversion that leads to radical changes in his or her perspective on an array of topics. Researchers also find that repeated exposure to a stimulus object is sufficient to enhance a person's opinion of the object. This theory underlies strategies used in advertising and marketing and can be seen at work, for example, in political campaigns in which greater candidate name recognition increases the probability that a candidate will win the election.

Although a person's opinions depend much on his or her personal experience, the source of many of these experiences is through parental teaching and modeling of parents' behavior. Gender role and racial opinions are two prominent areas where parental influence is often observed. Researchers have found that children's opinions toward gender are strongly influenced by their parents. Mothers who worked outside the home, for example, tend to have daughters who hold less traditional views toward work roles than those raised by mothers who were housewives. Researchers have also found that by the age of 4, many children tend to hold opinions that reflect cultural stereotypes about race, and that children tend to adopt their parents' racial prejudices. Parental influence on opinion formation has also been found to be significant in the development of political attitudes or political socialization, although as a child ages this influence wanes.

Other important influences on opinions are through institutions and social groups with which a person has contact, for example, schools, peers, and other reference groups. Schools tend to have an indoctrinating effect by disseminating not only knowledge required to formulate an opinion but also culture and values that constrain those opinions. For example, children may develop favorable feelings toward the president or the police as a result of classroom experiences. Peer groups tend to have a reinforcing effect on one's opinions. Voters, for example, often cite that they spoke with people in their social network about their vote decision and that their decision is similar to those of their friends. Finally, an individual's opinions can be influenced by the opinions of others in groups to which he or she belongs. However, this effect depends on two factors: (1) the degree of closeness a person feels to the group, and (2) the importance that person places on membership within the context of other groups that he or she belongs to.

Elite discourse (i.e., idea exchange among politicians, journalists, scholars, etc.) can also influence opinions. During periods of elite agreement, individuals tend to receive one-sided information on the topic, therefore increasing the likelihood that public opinion will reflect elite consensus. During elite disagreement, the individual receives two-sided information, leading to more varied opinions among the public. When the elite discourse is two-sided, people may attend to and be influenced by messages from elites whom they like or trust and disregard messages from elites whom they dislike or distrust. The elite discourse during the Vietnam War has been cited as one example of this effect. During the early years of the war, most of the political elites agreed on a containment strategy in Vietnam, but in the later years, as casualties mounted, elites disagreed, and public support for the war declined substantially.

With several thousand television, radio, and print media outlets in the United States, the mass media also has a significant role in affecting opinions. The majority of research today supports indirect effects of the media—although the media provides information on a range of topics, this information often does not substantially change public attitudes. The media does, however, influence public opinion indirectly via agenda setting, whereby media coverage influences which issues are salient to the public.

A person's attentiveness to information about an issue during opinion formation and change and his or her subsequent knowledge about the issue can have a significant impact on the strength of his or her opinion on that issue. Researchers have found that people tend to form fickle opinions when they do not attend to information about the issue and have little information on the topic. In contrast, a person's opinion on a subject may be more stable if it is based on a large body of evidence.

Jamie Patrick Chandler

See also Agenda Setting; Attitudes; Issue Definition (Framing); Opinion Norms; Opinion Question; Public Opinion

Further Readings

Asher, H. (2007). *Polling and the public* (7th ed.). Washington, DC: CQ Press.

Erikson, R. S., & Tedin, K. L. (2007). *American public opinion: Its origins, content, and impact*. New York: Pearson Longman.

Oskamp, S. P., & Schultz, W. (2004). *Attitudes and opinions*. Mahwah, NJ: Lawrence Erlbaum.

Taber, C. S. (2003). Information processing and public opinion. In D. O. Sears et al. (Eds.), *Oxford handbook of political psychology* (pp. 433–476). New York: Oxford University Press.

Zaller, J. R. (1992). *The nature and origins of mass opinions*. New York: Cambridge University Press.

Optimal Allocation

Optimal allocation is a procedure for dividing the sample among the strata in a stratified sample survey. The allocation procedure is called "optimal" because in a particular survey sampling design (stratified simple random sampling) it produces the smallest variance for estimating a population mean and total (using the standard stratified estimator) given a fixed budget or sample size.

A sample survey collects data from a population in order to estimate population characteristics. A stratified sample selects separate samples from subgroups (called "strata") of the population and can often increase the accuracy of survey results. In order to implement stratified sampling, it is necessary to be able to divide the population at least implicitly into strata before sampling. Given a budget that allows gathering data on n subjects or a budget amount $\$B$, there is a need to decide how to allocate the resources for data gathering to the strata. Three factors typically affect the distribution of resources to the strata: (1) the population size, (2) the variability of values, and (3) the data collection per unit cost in the strata. One also can have special interest in characteristics of some particular strata that could affect allocations.

In a stratified simple random sample, a sample of size n_h is selected from strata or subpopulation h, which has a population size of N_h $(h = 1, 2, \ldots, H)$. The standard estimator of the population total is $\sum_{h=1}^{H} N_h \bar{y}_h$, where $\bar{y}_h$ is the mean (arithmetic average) of the sample values in stratum h and $\sum$ denotes summation across strata $h = 1, 2, \ldots, H$. The variance of the estimator is $\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$, where S_h^2 is the variance of the values in stratum h. If the rate of sampling is small in all strata, then (ignoring the finite population correction terms $\left(1 - \frac{n_h}{N_h}\right)$) the variance is approximately $\sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h}$. Suppose the cost to collect data from one element (person, unit, etc.) in stratum h is C_h. If there is a budget of B, then the entire budget is spent when $B = \sum_{h=1}^{H} n_h C_h$. Then the variance (ignoring the finite population correction terms) of the estimated population total is minimized when the sample size in stratum h is $n_h = nN_h S_h / \sqrt{C_h} / \sum_{g=1}^{H} N_g S_g / \sqrt{C_g}$, where the summation in the denominator is over all strata, S_h is the standard deviation (square root of the variance) of the values in stratum h, and n is the total sample size. This formula implies that one should sample more in large subpopulations (strata), more in strata with large variances, and more in strata with small cost. If costs of per unit data collection are the same in all strata, then the optimal allocation in stratum h is $n_h = nN_h S_h / \sum_{g=1}^{H} N_g S_g$. If in addition variances (and standard deviations) are constant, then $n_h = nN_h / \sum_{g=1}^{H} N_g$, which is the allocation known as *proportional allocation* to strata. If the n_h's are not integers, then one must round the numbers to integers for sample selection. Rounding does not necessarily move all values to the closest integer for all strata, because the total sample size n needs to be allocated.

Suppose one wanted to collect data on students at a large public university. Questions of interest could be hours worked per week; amount of money expended per semester on textbooks; amount of time spent eating at restaurants in a week; number of trips to the airport in a semester; and whether or not friends smoke cigarettes. The students selected for the survey could be contacted via their university email addresses and asked to complete an online Web survey. A survey can be preferable to contacting every student, because for a sample better efforts can often made to encourage response and check data quality. Administrative records contain college year designations (first, second, third, fourth) for each student in the target population; college years can be used as strata. Suppose the total sample size is allowed to be 1,600 students. Equal allocation to strata would sample 400 students from each year. Table 1 presents allocations of students to the four strata based on total enrollments by college year; these numbers are similar to 2006 enrollment at Iowa State University. The hypothetical variable being considered is hours worked per week. It is assumed that students in higher years have more variable employment situations than students in earlier years, hence the increasing standard deviation. It also is assumed that more attempts are needed to contact students in later years than in earlier years. As can be seen in the table, the stratum of fourth-year students receives the largest sample ($n_4 = 731$), whereas the stratum of first-year students receives the smallest ($n_1 = 224$).

Table 1 Optimal allocation of 1,600 students to four strata defined by college year

Year	*Population Size: Total Enrollment*	*Standard Deviation: S_h*	*Per Unit Data Collection Cost: C_h*	*Sample Size:* $n_h = \frac{nN_h S_h/\sqrt{C_h}}{\sum_{g=1}^{H} N_g S_g/\sqrt{C_g}}$	*Sample Size: Rounded Values*
First	5,000	2.9 hours	1.0	223.8	224
Second	4,058	4.4 hours	1.1	262.8	263
Third	4,677	5.8 hours	1.2	382.3	382
Fourth	6,296	8.9 hours	1.4	731.1	731
Total	20,031			1,600.0	1,600

If, instead of being unequal, the costs per stratum were constant, then even more sample would be allocated to the strata with more advanced students. The first- through fourth-year strata would receive 201, 247, 376, and 776 students, respectively. If one planned to report on the four strata separately, then some compromise between optimal allocation and equal allocation would be advisable so that sample sizes in all strata remain reasonably large.

In practice, the costs associated with data collection and the variances (and standard deviations) in the various strata are unlikely to be known exactly. Instead, a preliminary small survey (a *pilot survey*) might be used to test methods of data collection and to collect data in order to estimate information needed for more efficient allocation. Alternatively, information that has been published or that is available in administrative records might be useful, or past experience in other surveys might be used.

Optimal allocations determined by one outcome variable, such as hours worked per week, are not necessarily the same as those determined using a different outcome variable, such as grade point average or body mass index. If several variables are considered important outcomes for a survey, then one could determine optimal allocations for each based on available data and select a compromise among the various allocations. It is likely that the average or compromise among allocations will be closer to proportional allocation than the most extreme allocations determined by optimal allocations based on the several variables.

Michael D. Larsen

See also Disproportionate Allocation to Strata; Elements; Finite Population Correction (fpc) Factor; Proportional Allocation; Strata; Stratified Sampling

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Ordinal Measure

Within the context of survey research, *measurement* refers to the process of assigning values to characteristics of individuals to indicate their position on an underlying construct, such as their level of satisfaction with the government or their political party affiliations. Ordinal measures are used to produce ordered rankings among values. For example, measurements or responses to the question, *In general, would you say your health is: excellent, very good, good, fair, or poor?* can be sorted and ordered from healthiest ("excellent") to least healthy ("poor"). Ordinal measures convey information about the relationship between values—that one value is greater than another—but they do not indicate how much greater a value is. Although "excellent" is greater in value than "very good," one cannot say with certainty that the distance between those two values is the same, less, or more than the distance between "very good" and "good."

Of the four levels of measurement, ordinal measures are more sophisticated than nominal measures but less statistically powerful than interval or ratio measures. With nominal measures (e.g., political party identification), numbers may be assigned arbitrarily to categories to distinguish among them, as the numbers themselves do not have an inherent value. With interval measures (e.g., IQ scores), the distance between values is equivalent, but unlike ratio-level measures (e.g., age), they do not include a true zero as a value. Characteristics of scales generally determine the appropriate statistics. Ordinal scales are best suited for nonparametric statistics such as modes and chi-square, but they often also are used for correlations, analyses of variance, and in mathematical models. Technically, means are not meaningful measures because of the categorical nature of ordinal data; that is, medians should be used as central measures of tendency. However, means and other statistics appropriate for interval data are used by many researchers willing to accept the uncertain differences between ordinal ranks. Ordinal scales with fewer than 5 points should probably not be treated as interval level, because the small number of data points may mask large differences between scale values.

Ordinal measures are typically obtained with ordinal scales that include closed-ended response categories in which the categories are labeled using words, numbers, or some combination of both. Key decisions in obtaining ordinal measures include how many categories or scale points to administer and how to label the points. Ordinal scales typically range from 3 to 11 points (e.g., 0–10 scale). In general, data quality is higher when measured using 5 to 7 points. A guiding principle in constructing ordinal scales is to develop categories that are balanced and approximately equal distance from one another. Likert scales, a popular type of ordinal scale, demonstrate this balance well. Likert scales are bipolar and include categories with both positive and negative values. A typical example is one in which respondents are asked their level of agreement with a particular statement, with response options ranging from "strongly disagree," "somewhat disagree," "neither," "somewhat agree," to "strongly agree." With regard to labeling, decisions include whether to label all of the categories or just the end points with verbal descriptions, or whether to label the categories with a combination of verbal descriptions and numbers. Overall, data quality is optimized when every scale point is represented by a verbal description.

Jennifer Dykema, Steven Blixt, and John Stevenson

See also Balanced Question; Bipolar Scale; Closed-Ended Question; Interval Measure; Level of Measurement; Likert Scale; Nominal Measure; Ratio Measure

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Krosnick, J. A., & Fabrigar, L. R. (in press). *The handbook of questionnaire design*. New York: Oxford University Press.

Outbound Calling

Telephone calls involving call centers are classified as inbound or outbound, depending on whether the call is being received by the call center (inbound) or initiated in the call center (outbound).

Any telephone survey that starts with a list of telephone numbers involves outbound calling, although telephone surveys often use inbound calling in support functions and can also be entirely inbound.

The list of numbers used might be randomly generated, such as a random-digit dialed (RDD) sample, created from public listings such as the white pages or from prior respondent contact (for example, a customer satisfaction survey might use the phone numbers provided by the customer at the point of purchase).

Particularly with RDD surveys, only a small proportion of dials made will result in live connects, and therefore the efficiency of outbound calling can be significantly improved by using dialer technology. Functions of particular value to conducting outbound surveys include the following:

1. *Autodispositioning*, in which the outcome of certain types of calls (e.g., busy, fax, disconnected) can be automatically detected from the signal tones, saving interviewer time and increasing accuracy in the assignment of dispositions, and

2. *Autodialing*, where the act of dialing is performed automatically on some trigger, such as
   - A keystroke instruction from an interviewer
   - The interviewer logging into the system or completing an interview, or
   - In the case of predictive dialers, some combination of the probability of a dial being answered and the probability of an interviewer being free to handle a connected call

Outbound calling is often more successful when supported by pre-survey notification to the selected sample, such as by an advance letter sent to numbers in the sample that can be matched to an address. For some numbers that cannot be associated with any postal address, pre-survey notification is still possible using text messaging or pre-recorded voice messages (some dialers have the ability to automatically dial and leave a pre-recorded message on any line that answers), although there have been mixed findings on whether these nonletter forms of pre-notification help or harm response rates.

The use of caller ID is another feature for which there is a varying impact on response rates. Some data suggest that a well-known survey name embedded in the caller ID can help response, and it is known that some exchange systems and household systems will not receive calls that do not have a caller ID associated with them. Other data suggest that households are more likely to answer the phone on a refusal conversion call if the caller ID is suppressed or different from that used on earlier dials. Finally, most calls go through many exchanges between the call center from which the call originates and the target telephone number that can change or transform the ID transmitted, introducing additional uncertainty about the impact of caller ID on the success of an outbound call.

Jenny Kelly

See also Advance Contact; Advance Letter; Caller ID; Directory Sampling; Inbound Calling; List-Assisted

Sampling; Matched Number; Predictive Dialing; Random-Digit Dialing (RDD); Refusal Conversion

Further Readings

Kelly, J., Link, M., Petty, J., Hobson, K., & Stork, P. (2008). Establishing a new survey research call center. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 317–339). New York: Wiley.

Trussell, N., & Lavrakas, P. J. (2005, May). *Testing the impact of caller ID technology on response rates in a mixed mode survey*. Paper presented at the 60th annual American Association for Public Opinion Research Conference, Miami Beach, FL.

OUTLIERS

An *outlier*, as the term suggests, means an observation in a sample lying outside of the "bulk" of the sample data. For example, the value "87" is an outlier in the following distribution of numbers: 2, 5, 1, 7, 11, 9, 5, 6, 87, 4, 0, 9, 7. This original meaning has been expanded to include those observations that are influential in estimation of a population quantity. Influence of an observation in estimation is intuitively understood as the degree to which the presence or absence of that observation affects the estimate in terms of the variance.

The notion of outliers is common in all statistical disciplines. However, it has a distinctive meaning in sample surveys for mainly two reasons: (1) sample surveys mostly deal with finite populations, often without assuming a parametric distribution; and (2) sample surveys often employ complex sample designs with unequal inclusion probabilities. Moreover, the meaning and handling of outliers differ also, depending on the stage of the survey process at hand: sample design stage, editing stage, and estimation stage.

The occurrence of outliers is frequently unavoidable when a multi-purpose survey is conducted. It may be nearly impossible to make the design efficient for all variables of interest in a large-scale multi-purpose survey. The outliers that have the most impact come from sample units that have a large sample value coupled with a large sampling weight. Probability proportional to size sampling (PPS) or size stratification is often used in the design stage to prevent such a situation from occurring. If the measure of size (MOS) is not reliable, it is difficult to eliminate outliers entirely, unless a census or a sample survey with a high sampling rate is used. This problem is especially pronounced when dealing with a volatile population such as businesses in economic surveys. A typical situation is that a unit with a small MOS, and thus assigned a small probability, has grown to have a medium or large value at time of observation, resulting in a huge weighted value.

In the editing stage of a survey process, outlier detection is performed to find extreme values that may be due to some survey error (response error or keying error during data entry). Such outliers are detected by comparing individual data values with others using a standardized distance measure defined as the absolute difference of the value from the center of the data (location) divided by a dispersion measure (scale). Using the sample mean and sample standard deviation to define the distance tends to mask outliers. To avoid the *masking effect*, estimates that are robust (insensitive) to outliers should be used. For example, one may use the median to estimate location and either the interquartile range (w), the difference between the third quartile (i.e., 75th percentile) and the first quartile (i.e., 25th percentile), or the mean absolute difference—which is the median of observations' absolute differences from the sample median. Weighted values of these quantities rather than unweighted values may be used if the sampling weights are available. Once the standardized distance is defined, a criterion by which an outlier is detected is set as the *tolerance interval* of the standardized distances; if an observation falls outside of the interval, it is declared as an outlier. The interval can be symmetric, nonsymmetric, or one-sided, depending on the underlying population distribution.

In the estimation stage, the impact of outliers is evaluated and treated. The influence of an observation depends on the estimator, and so outliers that were detected in the editing stage may or may not influence the values of the estimator. There are basically three approaches for treating outliers: (1) trimming or discarding of outliers, (2) *Winsorization* of outliers (i.e., replacing outliers with the largest or smallest "inliers" (which fall in the tolerance interval), (3) down-weighting of outliers. Trimming is seldom used in survey sampling. Winsorization can be effective if the weights are equal or similar. A variant of Winsorization is sometimes applied using pre-determined cutoff values (e.g., tolerance interval boundaries) instead of using the observed values to replace the outliers. When Winsorization is applied to the weighted data, it defines a hybrid method that modifies the sample values and weights simultaneously in a systematic way. Down-weighting can be achieved in various ways. Sometimes, post-stratification is used to create a post-stratum of outliers and their weights are reduced. A robust estimation technique such as *M-estimation* can be applied, which automatically detects outliers and down-weights them. It is preferable to limit the modified sampling weights to no less than unity because each sample observation should represent at least itself.

From the design-based perspective, any treatment of outliers in estimation introduces some bias in exchange for reduction of the sampling variance, and thus it can be seen as a bias–variance trade-off. Therefore, the mean square error (MSE) criterion is a good guiding principle by which to choose an estimator. Based on this principle, some people try to define an estimator that minimizes estimated MSE. Bayesian methodology is sometimes employed to achieve a bias–variance trade-off. However, it should be noted that often it is notoriously difficult to estimate the MSE reliably.

The influence of an outlier in estimation depends on the sample size. If the sampling rate is large or the sample size is large, the problem of outliers may be less troublesome because the variability in the estimator will be small. However, even with a large sample, outliers can cause a problem for domains with small domain sample sizes, where estimators may be less stable (i.e., more variable). In a multi-purpose survey, a set of weights that is too volatile can create an outlier situation for any variable. To control the variability of weights without considering a particular variable, weight trimming is often used. It should not be confused with the trimming technique that involves discarding of observations. The weight trimming simply reduces extreme weights to an appropriately chosen cutoff to control the weight variability. The technique itself resembles Winsorization, and it may be better to call it "weight Winsorization" rather than "weight trimming."

Hyunshik Lee

See also Design-Based Estimation; Finite Population; Mean Square Error; Probability of Selection; Probability Proportional to Size (PPS) Sampling

Further Readings

Lee, H. (1995). Outliers in business surveys. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business survey methods* (pp. 503–526). New York: Wiley.

Out of Order

The out of order survey disposition is used in telephone surveys when the telephone number dialed by an interviewer is nonworking or not in service. Although each local telephone company in the United States handles out-of-order telephone numbers differently, most companies in urban and suburban areas include a recording on these numbers that says something like, "The number you have dialed is a nonworking number," or "The number you dialed is not in service at this time." Some telephone companies in rural areas also include these recordings as standard practice. Thus, interviewers dialing out-of-order numbers in these areas may hear nothing, or the number may ring and ring, making it difficult to determine whether these numbers should be coded using the out of order disposition or as a ring–no answer noncontact.

In most telephone surveys, a case with an out of order disposition would be considered ineligible. Out of order dispositions are considered final dispositions and typically are not redialed again during the field period of a survey.

One other exception to these rules occurs if a telephone number in the sampling pool is temporarily disconnected or temporarily out of service, or if the number has a recording that indicates that the line is being checked for trouble. If there is clear evidence that a number in the sampling pool is temporarily out of order, this number should not be considered ineligible but instead should be considered a case of unknown eligibility. Assuming the field period permits it, most survey organizations will wait a few days and attempt to redial numbers that appear to be temporarily out of service.

Matthew Courser

See also Final Dispositions; Missing Data; Noncontacts; Response Rates; Temporary Dispositions; Unknown Eligibility

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Out of Sample

The out of sample survey disposition is used in all types of surveys, regardless of mode. Cases with out of sample dispositions are considered ineligible and are not contacted again. As a result, the out of sample disposition is considered a final disposition.

In a telephone survey, out of sample dispositions usually occur when the telephone number dialed by an interviewer rings to a household, business, or individual that is outside of the geographic sampling area for a survey. For example, in a random-digit dial (RDD) survey of the general public, this is most common when the survey is sampling relatively small geographic areas such as counties, towns or villages, or neighborhoods for which telephone prefix boundaries do not conform exactly (or even closely) to geopolitical boundaries. Out-of-sample cases usually are discovered only if the questionnaire for the telephone survey includes screening questions that verify that the respondent or household is located within the geographic area being sampled for the survey.

In an in-person survey, out-of-sample cases include ineligible housing units that were listed as being within the sampling area but actually are outside of it. Out-of-sample cases in these surveys also can include other households or businesses that were incorrectly included in a list sample—any unit that is not properly part of the target population for the survey.

In a mail or Internet survey of named respondents, out-of-sample cases occur when the named respondent is found to be ineligible to participate in the survey based on screening information he or she provides on the questionnaire. For example, a respondent who indicates that he or she is not a doctor would be considered out of sample in a mail or Internet survey of physicians.

Matthew Courser

See also Final Dispositions; Ineligible; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Overcoverage

Overcoverage occurs in survey sample frames when the frame contains more than enough sample records. This primarily results from two situations. In one case, there are records in the sample frame that do not contain respondents or members of the target population. In other cases, the same respondent is targeted by duplicate or multiple records in the sample frame. In either case, the sample frame contains sample records that should be interviewed.

Different types of overcoverage are commonly referred to as "ineligible units" or "multiple records." Different researchers use the term *overcoverage* inconsistently, so it is important to consider whether overcoverage in a given sample frame is caused by ineligible units, multiple records, or both.

Sample frames ideally contain a perfect one-to-one correspondence between sample records and members of the target population for a survey. In some cases, multiple records refer back to a single member of the target population. This type of overcoverage is sometimes called a "multiplicity of elements." In other cases, sample records fail to lead to members of the target population. These cases are sometimes referred to as "blanks" or "foreign elements."

Multiple sample records that refer to a single member of the target population are common in sample frames. In cases in which directories or lists are used as sampling frames, respondents can be included more than once if lists are compiled from multiple sources. More commonly, multiple records lead back to a single respondent when the sample frame and target populations are measured (covered) inconsistently. For example, if telephone numbers are sampled for a survey of households, a household with multiple telephones will be included multiple times. If sales records are used as a source for a consumer survey, then consumers who have made multiple purchases might be included in a sample multiple times.

Overcoverage caused by duplicate or multiple records can be adjusted for either by cleaning the sample frame or by providing sample weights to adjust for different probabilities that a respondent is included in the sample frame. Frame cleaning can be accomplished either before or during the survey field process. Cleaning before involves checking the sample frame for duplicate or multiple records and eliminating them. This "de-duping" is a basic part of constructing and checking a sample frame and is made enormously easier and more practicable by increased computer power.

The second type of overcoverage, in which sample records do not contain valid members of the target population, is present in almost all sample frames. This occurs for a variety of reasons. In some cases, sample records do not correspond to anything similar to the target population. For example, telephone samples often contain disconnected telephone numbers or numbers that have not been assigned. Household surveys may send an interviewer to an empty lot. A business directory might contain mailing addresses for establishments that went out of business many years ago. These listings are referred to as "blanks," "empty records," "empty listings," or more colloquially as "bad records" or "duds." In other cases, the sample record reaches a unit that can be screened for eligibility, but the record turns out to not be a member of the target population for the survey. For example, a survey of eligible voters may reach nonvoters, a survey that targets telephone households in one city instead may reach some households in a neighboring town, or a survey of college students may reach some recent graduates. These records are called "foreign elements," "out-of-scope units," or, colloquially, "screen-outs."

Survey researchers attempt to identify blanks or foreign elements by screening to determine whether they are eligible for the survey. In some cases, this can be done automatically. For example, residential telephone samples can be screened against databases of known business households, and other electronic matching can identify other foreign elements. In many cases, however, an interviewer or other field staff member needs to contact a sample record to determine if it is an eligible member of the survey's target population. This is especially true for studies that utilize general population sample frames to identify rare subpopulations. For example, a survey of parents with disabled children who live at home may need to contact and screen all households in the sample to locate the eligible households, even though the majority of households do not have children living there and most others have children who are not disabled. These low-incidence samples can add great cost to a survey.

Blanks and foreign elements generally do not lead to biased or distorted survey results, but they often result in a loss of both sample and economic efficiency. Surveys that desire a specific degree of statistical precision need to increase (inflate) initial sample sizes to account for these records. More important for many researchers, the cost implications of conducting surveys that contain many ineligible units make interviews with many rare or low-incidence populations impracticable.

Chase H. Harrison

See also Coverage Error; Duplication; Eligibility; Sampling Frame; Survey Costs

Further Readings

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.
Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

Overreporting

In many surveys, respondents tend to report more socially desired behaviors than they actually performed. In addition to this type of misreporting—called "overreporting"—respondents are also inclined to understate that they have engaged in socially undesirable behaviors, which is called "underreporting." Similar to underreporting, overreporting is assumed to be connected to social desirability bias and thus occurs on the cognitive editing stage of the question–answer process.

Among other topics, overreporting of voting and being registered to vote has been in the focus of methodological research for decades. Since respondents in national- and state-level and local election polls tend to overly state that they have voted in the election, voter turnout has traditionally been overestimated. Usually, overreporting is identified applying post-survey validations using record checks (like in the National Election Study).

Since not every survey can afford a cost-intensive validation study, several attempts have been made in order to reduce vote overreporting, either by softening the question wording so that respondents will not feel embarrassed to admit that they have not voted or by a set of preceding questions on voting behavior in other, prior elections. It was assumed that respondents would be more willing to admit that they have not voted in the most recent past election if they were able to report voting in previous elections. However, neither tactic succeeded—the proportion of vote overreporting remained unaffected.

Overreporting is associated with respondent characteristics. Respondents who hold strong positive opinions on a particular behavior are more likely to falsely report this behavior in a survey.

Marek Fuchs

See also Cognitive Aspects of Survey Methodology (CASM); Errors of Commission; Record Check; Respondent-Related Error; Sensitive Topics; Social Desirability; Underreporting

Further Readings

Belli, R. F., Traugott, M. W., Young, M., & McGonagle, K. (1999). Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring. *Public Opinion Quarterly*, *63*(1), 90–108.

Presser, S. (1990). Can changes in context reduce vote overreporting in surveys? *Public Opinion Quarterly*, *54*(4), 586–593.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

P

Paired Comparison Technique

The paired comparison technique is a research design that yields interval-level scaled scores that are created from ratings made by each respondent for all possible pairs of items under consideration. The basis for the method dates back to its first reported use in the mid-1800s. Although the technique is a very powerful approach for producing a highly reliable ranking of the rated items, it is underutilized by survey researchers due to the amount of data that often must be gathered, and thus its cost and the burden it places on respondents.

At the simplest level, paired comparisons (i.e., simultaneously comparing two things with each other) are made by each respondent among a set of items using a binary scale that indicates which of the two choices are most preferred, most pleasant, most attractive, or whatever other judgment the respondent is asked to make in comparing the two. However, more complex judgments can be generated by having respondents indicate their choices along a continuum of response choices rather than a simply binary choice (A or B).

For example, if a political pollster wanted to determine the relative ordering of voter preferences among five Republican primary candidates, a paired comparison design would yield the most valid data. In this design, each candidate would be paired with each of the other candidates, and each respondent would judge each pair on some criterion. Typically this would be done by using a scaled response format such as *Strongly Prefer Candidate A; Prefer Candidate A; Slightly Prefer Candidate A; Slightly Prefer Candidate B; Prefer Candidate B; Strongly Prefer Candidate B*. Generally the midpoint of the preference scale—which in this example would be, "Prefer Neither Candidate A nor Candidate B"—is not offered to respondents because it is reasoned that the likelihood that there is complete indifference between the two is extremely low. Providing this "no preference" choice may encourage some respondents to satisfice and use the middle option too readily.

Scoring using paired comparison data is straightforward. In the previous example a "Strongly Preferred" response would be scored with a 3, a "Preferred" response would scored with a 2, and a "Slightly Preferred" response would be scored with a 1. If Candidate A were paired with Candidate D, and Candidate A were "strongly preferred" over Candidate D by a given respondent, then the respondent would be assigned a $+3$ score for Candidate A for that pairing, and the respondent would get a –3 score for Candidate D for that pairing. The scaled scores for a specific candidate for each respondent would be the sum of the respondent's individual scores from each of the pairings in which that candidate was included. In the example of five candidates being paired in all possible ways, there would be $((c(c-1))/2)$ possible paired comparisons, with c representing the number of things being paired. Thus in this example there are $((5(5-1))/2)$ or 10 pairs: AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE. (The pairings would be presented in a random order to respondents.) Each pairing would require a separate question in the survey; thus, this five-candidate comparison would require 10 questions being asked of each respondent. If one of the candidates in this example were "strongly preferred" by a specific respondent over each of the other four candidates she or he was paired with, that candidate would get a score of +12 for this respondent. If a candidate were so disliked that every time she or he was paired with one of the other four candidates a given respondent always chose "Strongly Preferred" for the other candidates, then the strongly disliked candidate would be assigned a scaled score of –12 for that respondent. Computing scale scores for each thing that is being rated is easy to do with a computer, and these scaled scores provide very reliable indications of the relative preferences a respondent has among the different items being compared. That is, asking a respondent to rank all of the things being compared in one fell swoop (i.e., with one survey question) will yield less reliable and valid data than using a paired comparison design to generate the ranking.

Unfortunately, increasing the numbers of things being compared in a paired comparison design quickly causes many more pairings to be required for judgment by each respondent. Thus, if a researcher wanted 10 things to be compared, a total of 45 pairings would need to be judged by each respondent. In instances when a large number of pairings are to be made by respondents, a researcher is wise to add some reliability checks into the set of pairings. This is done by randomly selecting some of the pairs and reversing the order of the things within those pairings. For example, if the AF pairing were randomly chosen as one of the reliability checks, and if A was compared with F earlier in the question sequence by the respondent, then later on in the sequence F would be compared with A. The respondent is not likely to recall that she or he had already made this comparison if the set of items being compared is large. Adding such reliability checks allows the researcher to identify those respondents who are not taking the task seriously and instead are answering haphazardly.

Paul J. Lavrakas

See also Interval Measure; Ranking; Rating; Respondent Burden; Satisficing

Further Readings

Brown, T. C., Peterson, G. L., & Brink, G. E. (2007). *An enquiry into the method of paired comparison: Reliability, scaling, and Thurstone's law of comparative judgment.* Fort Collins, CO: Rocky Mountain Research Station, U.S. Forest Service. Retrieved June 4, 2008, from http://www.fs.fed.us/rm/value/brown.html

David, H. A. (1988). *The method of paired comparisons* (2nd ed., Vol. 41). New York: Oxford University Press.

Lavrakas, P. J. (1975). Female preferences for male physiques. *Journal of Research in Personality*, *9*, 324–334.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychology Review*, *34*, 273–286.

Thurstone, L. L. (1927). The method of paired comparisons for social values. *Journal of Abnormal Social Psychology*, *21*, 384–400.

Panel

A panel refers to a survey sample in which the same units or respondents are surveyed or interviewed on two or more occasions (*waves*). Panels can give information about trends or changes in the characteristics of a given population over a period of time. A panel usually can measure changes in the characteristics of interest with greater precision than a series of independent samples of comparable size. A survey using a panel design is often called a "longitudinal survey," which is one particular type of repeated survey.

The sample design for a panel is very different from the one for an independent sample or a series of independent samples. In the sample design for a panel, more stable stratification variables over time can (and should) be employed than when using independent samples, because whereas a panel design may be statistically efficient in a short run, it may not be over a longer period of time. Also, the design for a panel should incorporate the changes in the population that the panel is meant to represent, such as births and other additions and deaths and other removals of sample units, in an optimal way so as not to cause disruption to the ongoing survey operations at different points of time or waves.

Advantages and Disadvantages

There are clear advantages to using a panel rather than using a series of independent samples in studying a target population. Some of these include the following:

1. A panel provides the details on the nature of change. For example, suppose that from one independent sample to another the prevalence of a disease changes from 5% to 15%. We know the

simple change of 10 percentage points, but do not know whether the incidence of new cases is 0% or 10% or something in between 0% and 10%. Using a panel, the percentage of new cases is easily obtained.

2. It reduces the variability for estimates of change at different points in time. For example, if P_1 and P_2 are the estimated proportion of unemployed people at Time 1 and Time 2, respectively, then the estimate of change is $P_2 - P_1$. When using a panel, the variance of $P_2 - P_1$ is reduced relative to the one for two independent samples.
3. It often reduces the observational errors by interviewers or respondents. For example, as the surveys are repeated, interviewers have a better experience of administering the interviews and respondents have a better understanding of the questionnaire.
4. It gives information on the dynamic behavior of respondents over time. For example, it is possible to explain the fact that the same people experiencing an event in the past, say unemployment, tend to experience it in the future.

However, there also are disadvantages to using a panel:

1. Analysis of a panel data is more complicated due to differential unit and item nonresponse and wave nonresponse, as well as the complexity of sample design. *Wave nonresponse* occurs when one or more waves of panel data are missing for a sample unit that has responded for at least one wave. Weighting, imputation, or a combination of weighting and imputation can be employed to compensate for missing wave data.
2. Measuring changes in individuals from one year to another year may be unreliable because reluctant respondents may give poor answers to repeated interviews, or respondents may refuse to be interviewed several times due to panel fatigue, resulting in higher nonresponse over time.
3. Respondents' answers to questions in previous waves may affect their answers in subsequent waves; this is termed *panel conditioning*.
4. It is difficult to keep the panel representative during a long period of time because the target population can change over time.
5. It may be too expensive to locate all respondents a year later or after a certain period of time, due to travel costs and the obstacles to following and finding some respondents who have moved without any new contact information.
6. It can be difficult to identify the same sample units over time. For example, identification of the same family units can be complicated when the family composition is changed by births, deaths, marriages, divorces, and so on.

Rotating Designs

Even if a panel is a "bad" sample, that is, if it does not well represent a given population over time, the organization carrying out the panel survey may have to continue to maintain that panel. But there is a solution to such a problem of the panel. It is a *rotating panel design*, in which a part of the panel sample is replaced at each subsequent point in time. This design is intermediate (i.e., a hybrid) between a panel sample and independent samples. As the simplest example, one may choose a panel design involving overlaps of half of the sample elements, as shown by AB, BC, CD, DE, EF, and so on at different points of time. In this example, the B panel sample appears in the first and second waves of data collection; the C panel sample appears in the second and third waves, and so on. Such rotating designs not only reduce respondent burden but also provide an opportunity to refresh the sample at each point of time with cases that better reflect the current makeup of the target population.

Examples

A number of panel surveys with economic or social science focus have been conducted around the world. One of the oldest panel surveys is the Panel Study of Income Dynamics (PSID) conducted by the Survey Research Center at the University of Michigan. This panel study has collected information on the dynamic aspects of economic and demographic behavior, including sociological and psychological measures. The panel of the PSID is a representative sample of U.S. family units and individuals (men, women, and children). The panel, originating in 1968, consisted of two independent samples: a cross-sectional national sample and a national sample of low-income families. The cross-sectional sample, which yielded about 3,000 completed interviews, was an equal probability sample of households from the 48 contiguous states. The national sample of low-income families came from the Survey of Economic Opportunity (SEO) conducted by the U.S. Bureau of the Census. This second sample consisted of about 2,000 families from SEO respondents with heads under the age of 60. The original core sample combined by the two samples was increased to approximately 6,168 in 1997 and nearly 7,400 in 2005. These changes were to reflect the changing nature of immigration in the United States. The PSID was collected in face-to-face interviews using paper-and-pencil questionnaires between 1968 and 1972. Thereafter, the majority of interviews were conducted by telephone, and in the 1999 wave, 97.5% of the interviews were conducted by computer-assisted telephone interviewing.

Sun Woong Kim

See also Attrition; Item Nonresponse; Longitudinal Studies; Panel Conditioning; Panel Data Analysis; Panel Survey; Respondent Burden; Rotating Panel Design; Wave

Further Readings

Hill, M. S. (1992). *The panel study of income dynamics: A user's guide.* Newbury Park, CA: Sage.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. (Eds.). (1989). *Panel surveys.* New York: Wiley.

Kish, L. (1995). *Survey sampling.* New York: Wiley.

Panel Conditioning

Panel conditioning is an effect sometimes observed in repeated surveys when a sample unit's response is influenced by prior interviews or contacts. Various possibilities have been suggested to explain the cause. Panel conditioning can affect the resulting estimates by introducing what is sometimes called "time-in-sample bias" or "rotation group bias."

In many surveys, the household, business, or other sample unit is contacted more than once over a period of time, usually to reduce the total survey cost, to produce longitudinal estimates, or to decrease the standard error of the estimate of change in the items of interest. In various documented studies, the levels of unemployed persons, expenditures, illness, victimizations, house repairs, and other characteristics were significantly higher or lower in earlier survey contacts than in later ones.

Potential scenarios to explain this behavior are extensive. At times a respondent recalls the answer to a question from a prior interview and repeats it, even when there is a change. Respondents can learn from their past experience answering a questionnaire. In some surveys, certain responses—for example, receiving some type of income or being a victim of a crime—may elicit a lengthy set of follow-up questions or probes. Over time, a respondent may observe this tendency and adjust her or his response to avoid being asked the follow-up sequence. On the other hand, the concepts or questions in a questionnaire may become clearer to a respondent after one or more contacts, producing responses based on better understanding or recall. In these instances, the conditioning can lead to more accurate data.

In surveys that ask for opinions, attitudes, or projected behavior, the person in a sample may become more informed or aware of the issues through a series of interviews. This can affect a later outcome by causing the respondent to explore the topic before the next interview or to change his or her behavior, for example, to vote for a specific candidate or simply to vote.

The effects of panel conditioning are not always initiated by the respondent. The procedures for conducting a repeated survey can differ from one interview to the next. For example, when the household or business is first contacted, additional relevant questions might be asked that are omitted in later interviews. This omission can influence the subsequent responses. Further, the interviewer may have access to responses from prior interviews and may change the way he or she conducts the next interview based on this information.

In a repeated survey, the effects of panel conditioning on the estimates are difficult to measure and correct for, in part because the effects may be confounded with actual change, panel attrition, the mode of data collection, or other factors. One way to study the effects is to operate a repeated panel simultaneously with independent cross-sections of the same population and compare the results at fixed points in time. Another approach is to compare responses to reliable administrative records and gauge the accuracy over the life of the panel.

Statistically, the group or panel of sample units responding for the first time will exhibit a bias if the mean of their responses differs from the true value based on the entire population. In the same way, the estimate from the panel responding for the second or third time can suffer from bias of a different value. Often this time-in-sample bias is measured by comparing a panel's value to the average over all panels, with the latter as a proxy for the "truth." However, without additional studies, the true bias of each panel cannot be determined; it can only be expressed relative to a number such as the average.

The relative effects among the panels may be studied under a balanced rotating panel design, in which the set of times in sample is the same in every period. Under such a design, if the time-in-sample biases are additive and remain constant over time, the biases can cancel each other relative to the truth when estimating change over time.

Patrick J. Cantwell

See also Attrition; Panel; Panel Fatigue; Panel Survey; Reinterview; Response Bias; Rotating Panel Design

Further Readings

Bailar, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, *70*, 23–30.

Kasprzyk, D., Kalton, G., Duncan, G., & Singh, M. P. (Eds.). (1989). *Panel surveys.* New York: Wiley.

Panel Data Analysis

Panel data analysis refers to the statistical analysis of data sets consisting of multiple observations on each sampling unit. This could be generated by pooling time-series observations across a variety of cross-sectional units, including countries, firms, or randomly sampled individuals or households. This also encompasses longitudinal data analysis in which the primary focus is on individual histories. Two well-known examples of U.S. panel data are the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys of Labor Market Experience (NLS). European panels include the German Socio-Economic Panel, the British Household Panel Survey, and the European Community household panel.

Benefits and Limitations

Some of the benefits and limitations of using panel data for statistical analysis include a much larger data set, because panel data are multiple observations on the same individual. This means that there will be more variability and less collinearity among the variables than is typical of cross-section or time-series data. For example, in a demand equation for a given good (say, gasoline), price and income may be highly correlated for annual time-series observations for a given country or state. By stacking or pooling these observations across different countries or states, the variation in the data is increased and collinearity is reduced. With additional and more informative data, one can get more reliable estimates and test more sophisticated behavioral models using less restrictive assumptions.

Another advantage of panel data is their ability to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to bias in the resulting estimates. For example, in an earnings equation, the wage of an individual is regressed on various individual attributes, such as education, experience, gender, race, and so on. But the error term may still include unobserved individual characteristics, such as ability, which is correlated with some of the regressors, such as education. Cross-sectional studies attempt to control for this unobserved ability by collecting hard-to-get data on twins. However, using individual panel data, one can, for example, difference the data over time and eliminate the unobserved individual invariant ability. Panel data sets are better able to study complex issues of dynamic behavior. For example, with cross-section data, one can estimate the rate of unemployment at a particular point in time. Repeated cross-sections can show how this proportion changes over time. Only panel data sets can estimate what proportion of those who are unemployed in one period remains unemployed in another period.

Limitations of panel data sets include problems in the design, data collection, and data management of panel surveys. These include the problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding to prevent the shifting of events from outside the recall period into the recall period, and time-in-sample bias.

Another limitation of panel data sets is distortion due to measurement errors. Measurement errors may arise because of faulty response due to unclear questions, memory errors, deliberate distortion of responses (e.g., prestige bias), inappropriate informants, misrecording of responses, and interviewer effects. Although these problems can occur in cross-section studies, they are aggravated in panel data studies. Panel data sets may also exhibit bias due to sample selection problems. For the initial wave of the panel, respondents may refuse to participate, or the interviewer may not find anybody at home. This may cause some bias in the inference drawn from this sample. Although this nonresponse can also occur in cross-section data sets, it is more serious with panels because subsequent waves of the panel are still subject to nonresponse. Respondents may die, move, or find that the cost of responding is high. The rate of attrition differs across panels and usually increases from one wave to the next, but the rate of increase typically declines over time.

Applications

Most panel data applications have been limited to a simple regression with error components disturbances, such as the following:

$$y_{it} = x_{it}'\beta + \mu_i + \nu_{it}, \quad i = 1, \ldots, n; \quad t = 1, \ldots, T,$$

where y_{it} may denote log(wage) for the ith individual at time t, and x_{it} is a vector of observations on k explanatory variables such as education, experience, race, sex, marital status, union membership, hours worked, and so on. In addition, β is a k vector of unknown coefficients, μ_i is an unobserved individual specific effect, and ν_{it} is a zero mean random disturbance with variance σ_ν^2. The error components disturbances follow a one-way analysis of variance (ANOVA). If μ_i denote fixed parameters to be estimated, this model is known as the *fixed-effects* (FE) model. The x_{it}'s are assumed independent of the ν_{it}'s for all i and t. Inference in this case is conditional on the particular n individuals observed. Estimation in this case amounts to including $(n-1)$ individual dummies to estimate these individual invariant effects. This leads to a large loss in degrees of freedom and attenuates the problem of multi-collinearity among the regressors. Furthermore, this may not be computationally feasible for large sample size panels. In this case, one can eliminate the μ_i s and estimate β by running least squares of $\tilde{y}_{it} = y_{it} - \bar{y}_{i.}$ on the $\tilde{x}_{it}$s similarly defined, where the dot indicates summation over that index and the bar denotes averaging. This transformation is known as the *within transformation*, and the corresponding estimator of β is called the *within estimator* or the *FE estimator*. Note that the FE estimator cannot estimate the effect of any time-invariant variable such as gender, race, or religion. These variables are wiped out by the within transformation. This is a major disadvantage if the effect of these variables on earnings is of interest.

If μ_i denotes independent random variables with zero mean and constant variance σ_μ^2, this model is known as the *random-effects* (RE) model. The preceding moments are conditional on the x_{it}s. In addition, μ_i and ν_{it} are assumed to be conditionally independent. The RE model can be estimated by generalized least squares (GLS), which can be obtained using a least squares regression of $y_{it}^* = y_{it} - \theta\bar{y}_{i.}$ on x_{it}^* similarly defined, where θ is a simple function of the variance components σ_μ^2 and σ_ν^2. The corresponding GLS estimator of β is known as the *RE estimator*. Note that for this RE model, one can estimate the effects of individual-invariant variables. The best quadratic unbiased (BQU) estimators of the variance components are ANOVA-type estimators based on the true disturbances, and these are minimum variance unbiased (MVU) under normality of the disturbances. One can obtain feasible estimates of the variance components by replacing the true disturbances by OLS or fixed-effects residuals.

A specification test based on the difference between the fixed- and random-effects estimators is known as the *Hausman test*. The null hypothesis is that the individual effects are not correlated with the x_{it}s. The basic idea behind this test is that the fixed-effects estimator $\tilde{\beta}_{FE}$ is consistent, whether or not the effects are correlated with the x_{it}s. This is true because the fixed-effects transformation described by $\tilde{y}_{it}$ wipes out the μ_i effects from the model. However, if the null hypothesis is true, the fixed-effects estimator is not efficient under the random-effects specification because it relies only on the within variation in the data. On the other hand, the random-effects estimator $\hat{\beta}_{RE}$ is efficient under the null hypothesis but is biased and inconsistent when the effects are correlated with the x_{it}s. The difference between these estimators $\hat{q} = \tilde{\beta}_{FE} - \hat{\beta}_{RE}$ tends to zero in probability limits under the null hypothesis and is nonzero under the alternative. The variance of this difference is equal to the difference in variances, $\mathrm{var}(\hat{q}) = \mathrm{var}(\tilde{\beta}_{FE}) - \mathrm{var}(\hat{\beta}_{RE})$ because $\mathrm{cov}(\hat{q}, \hat{\beta}_{RE}) = 0$ under the null hypothesis. Hausman's test statistic is based on $m = \hat{q}'[\mathrm{var}(\hat{q})]^{-1}\hat{q}$ and is asymptotically distributed a chi-square with k degrees of freedom under the null hypothesis.

Special Panel Data Sets

Space limitations do not allow discussion of panel data models that include treatment of missing observations, dynamics, measurement error, qualitative limited dependent variables, endogeneity, and nonstationarity of the regressors. Instead frequently encountered special panel data sets—namely, pseudo-panels and rotating panels—are discussed. *Pseudo-panels* refer to the construction of a panel from repeated cross-sections, especially in countries where panels do not exist but where independent surveys are available over time. The United Kingdom Family Expenditure Survey, for example, surveys about 7,000 households annually. These are independent surveys because it may be impossible to track the same household across surveys, as required in a genuine panel. Instead, one can track cohorts and estimate economic relationships based on cohort means. Pseudo-panels do not suffer the attrition problem that plagues genuine panels and may be available over longer time periods.

One important question is the optimal size of the cohort. A large number of cohorts will reduce the size of a specific cohort and the samples drawn from it. Alternatively, selecting few cohorts increases the accuracy of the sample cohort means, but it also reduces the effective sample size of the panel.

Rotating panels attempt to keep the same number of households in the survey by replacing the fraction of households that drop from the sample in each period with an equal number of freshly surveyed households. This is a necessity in surveys in which a high rate of attrition is expected from one period to the next. Rotating panels allow the researcher to rest for the existence of time-in-sample bias effects. These correspond to a significant change in response between the initial interview and a subsequent interview when one would expect the same response.

Other Considerations

Panel data are not a panacea and will not solve all the problems that a time-series or a cross-section study could not handle. For example, with macro-panels made up of a large number of countries over a long time period, econometric studies argued that panel data will yield more powerful unit root tests than individual time-series. This in turn should help shed more light on the purchasing power parity (PPP) and the growth convergence questions. This led to a flurry of empirical applications, along with some skeptics who argued that panel data did not save the PPP or the growth convergence problem.

Also, collecting panel data is quite costly, and there is always the question of how often one should interview respondents. For example, some economists argue that economic development is far from instantaneous, so that changes from one year to the next are probably "too noisy" (i.e., unreliable) and too short term to really be useful. They conclude that the payoff for panel data is over long time periods, 5 years, 10 years, or even longer. In contrast, for health and nutrition issues, especially those of children, one could argue the opposite case—that is, those panels with a shorter time span—are needed in order to monitor the health and development of these children.

Users of panel data argue that these data provide several advantages worth their cost. However, as with survey data in general, the more we have of it, the more we demand of it. The survey researcher using panel data, or any data for that matter, must know the data's strengths and limitations.

Badi H. Baltagi

See also Analysis of Variance (ANOVA); Attrition; Cross-Sectional Data; Longitudinal Studies; Measurement Error; Panel; Panel Conditioning; Panel Fatigue; Panel Survey; Repeated Cross-Sectional Design; Rotating Panel Design

Further Readings

Kalton, G., & Brick, M. (2000). Weighting in household panel surveys. In D. Rose (Ed.), *Researching social and economic change: The uses of household panel studies* (pp. 96–112). New York: Routledge.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (Eds.). (1989). *Panel surveys*. New York: Wiley.

Panel Fatigue

Panel fatigue refers to the phenomenon in survey research whereby the quality of data that is gathered from a particular member of a survey panel diminishes if she or he is expected to stay in the panel for too long a duration (i.e., for too many waves) of data collection. In the extreme, panel fatigue leads to premature panel nonresponse for particular panel members prior to their tenure in the panel officially expiring. That is, the respondent drops out of the panel early and thereafter is a source of *panel attrition*. Panel fatigue also contributes to item nonresponse (i.e., missing data), to increases in satisficing as a mode of response, and to other forms of lower quality of data. Because panel fatigue does not occur uniformly across all types of respondents, it often leads to *differential panel attrition*. Old adults and young adults, those with less educational attainment, and/or minorities are most likely to display panel fatigue.

The effects of panel fatigue are best countered by researchers making informed and reliable decisions about how long is "too long" for panel members to stay in a panel survey. These data quality considerations must be balanced with the cost implications of having to turn over (i.e., refresh) the panel more often than is desirable from a cost standpoint. For example, a conservative but expensive approach would be to determine when nonnegligible panel fatigue starts for the 20% or so of panel members who are the first to experience panel fatigue, and then limit all panel membership to that lower duration. Panel fatigue also can be countered by not timing waves of subsequent data collection too closely together; or by rotating random subsets of panel members in and out of data collection (e.g., every other wave, or every two of three waves).

Extra attention paid to panel members also may help counter panel fatigue. This includes friendly "staying in touch" communications from the researchers between waves of data collection that show sincere interest in the well-being of the panel members and subtly stress the importance of remaining active panel members. Use of contingent (performance-based) incentives also has been shown to be effective in reducing the negative effects of panel fatigue.

Paul J. Lavrakas

See also Attrition; Contingent Incentives; Missing Data; Panel; Panel Conditioning; Panel Survey; Rotating Panel Design; Satisficing; Wave

Further Readings

Holden, R., Heng, L. M., Lavrakas, P. J., Bell, S., & Flizik, A. (2006, May). *Personal contact and performance-based incentives to raise long-term panel compliance and to reduce missing data.* Paper presented at the 61st annual conference of the American Association for Public Opinion Research, Montreal, Quebec.

Kasprzyk, D., Duncan, G., Kalton, G., & Singh, M. P. (Eds.). (1989). *Panel surveys.* New York: Wiley.

Lansdowne, J. (1970). The mailed questionnaire in panel research: Some empirical observations. *Social Forces*, *49*(1), 136–140.

Panel Survey

The essential feature of a longitudinal survey design is that it provides repeated observations on a set of variables for the same sample units over time. The different types of longitudinal studies (e.g., retrospective studies, panel surveys, and record linkages) are distinguished by the different ways of deriving these repeated observations. In a panel survey, repeated observations are derived by following a sample of persons (a panel) over time and by collecting data from a sequence of interviews (or waves). These interviews are conducted at usually fixed occasions that in most cases are regularly spaced.

There are many variations under this general description of a panel survey, including (a) cohort panel surveys, (b) household panel surveys, and (c) rotating panel surveys. These three types of panel surveys can be distinguished, first, by the sampling units and the population the survey aims to represent. The focus can be entirely on individuals or on individuals within their household context. A second distinction is between surveys comprising a single panel of indefinite life and surveys comprising multiple overlapping panels of fixed life. Choosing the appropriate design for a panel survey depends on the priorities and goals of the (potential) data users and requires an assessment of the benefits of the different sorts of information collected and the costs required for deriving them.

Cohort Panel Surveys

A cohort panel survey is the simplest example of a single panel of indefinite life. It is an individual-level panel focusing on a population subgroup that has experienced the same event during the same time period (a cohort), such as having been born in a particular month, being a high school graduate in a given year, or having been married during the same year.

Cohort panel surveys are also called "fixed panel surveys," since the definition of membership of the cohort is fixed and cannot change over time. The rules for following the sample units in subsequent waves are simple: At each wave of the cohort panel survey, interviews are attempted with all original cohort members. After the initial sample selection, no additions to the sample are made.

Cohort panels are often set up to study long-term change and individual development processes, such as transitions into adulthood and marital or other union formation and dissolution. The data of the 1970 British Cohort Study (BCS70), for example, provide researchers with an opportunity to study the life-course experiences of a group of individuals representative of all men and women born in the 1970s in Great Britain. More specifically, the BCS70 follows more than 17,000 men and women born in Great Britain in a specific week in 1970. Since the first wave of data collection in 1970 (age 0), there have been six other major data collection waves, in 1975 (age 5), 1980 (age 10), 1986 (age 16), 1996 (age 26), 1999/2000 (age 29/30), and 2004/2005 (age 34/35). This cohort panel survey collected data on many aspects of the health, social development, and education of the cohort members as they passed through childhood and adolescence. In the more recent waves, the information collected covers transitions into adult life, including leaving full-time education, entering the labor market, setting up independent homes, forming partnerships, and becoming parents.

In cohort studies, individuals are sometimes followed through their entire life course, and because of this cohort studies often have longer times in between interviews or waves (e.g., 5 or 10 years for BCS70). However, a longer inter-interview period might result in a larger proportion of the sample that drops out because of panel attrition. This happens because the proportion of sample units who have moved increases for longer intervals and because the difficulties of tracking movers since the last wave become more severe when the intervals are longer. Some cohort panel surveys reduce some of the problems of tracing sample units by using more frequent interviewing. For example, the National Longitudinal Survey of Youth 1979 (NLSY79), a U.S. nationally representative sample of approximately 13,000 young men and women who were 14 to 22 years of age when first surveyed in 1979, interviewed cohort members annually through 1993 and bi-annually since 1994.

Household Panel Surveys

When the only objective of a panel survey is to facilitate longitudinal research at the individual level, it may be sufficient to adopt a cohort approach that simply follows the initial sample selected for the first wave. However, when cross-sectional population estimates for the life of the study are also important, it is necessary to update the sample at each wave to represent new entrants to the population of interest. The typical household panel survey has an indefinite life and is set up to study individual and household change; individuals are interviewed at fixed intervals, usually a year, and information is collected about them and the households in which they reside. The best-known household panel surveys may be the U.S. Panel Study of Income Dynamics (PSID), the British Household Panel Study (BHPS), and the German Socio-Economic Panel (*Das Sozio-oekonomische Panel*—SOEP). The PSID is considered the "grandfather" of all modern household panel surveys, as it originated in 1968. After a decade and a half of experience gained from the PSID, the German SOEP was established in 1984, followed by the British BHPS in 1991.

The main difficulty with household panel surveys is thus that they require a more complicated design to remain representative across time for both the individuals and the households in which they reside. The composition of nearly every population of interest, whether of individuals or households, changes over time. Individuals enter the population when they are born, immigrate or attain the age or another status that is used to define the population of interest, and depart when they die, emigrate, move to institutions, such as a home for the elderly, or in some other way lose that eligibility status. At the same time, families are "born" when children leave their parents and set up their own independent households or when a divorce or separation breaks up a family into two, and families "die" when all members of the original household die or when two households are merged into one through marriage or other living arrangements. In a household panel survey, in addition to making decisions about the sample design for the initial wave, important decisions need to be made about which sample units are to be retained in the panel at each wave to remain representative of a population that changes composition as a consequence of birth, death, and mobility during the life of the panel.

Following-rules are those rules that are designed to follow up and to update the initial sample, so as to ensure that on every wave throughout the survey's time span, the sample remains cross-sectionally representative of the population of interest. Following-rules are thus used to add new members to the sample and to remove members from the sample in the same way as persons are added to or removed from households in the broader population. In most household panels, all adults and children in the representative sample of households in the first wave are defined as *original* or *continuous sample members* (OSMs). A child born to or adopted by an OSM also counts as an original sample member. Temporary sample members (TSMs) are individuals who join the household of an original sample member after the initial wave. For example, a new partner who moves in with an OSM, or an elderly person who becomes dependent and moves in with a family member who is an OSM, would be considered a TSM. Most household panel surveys have adopted the rule that, at the second and subsequent waves, attempts are made to interview all adult members of all households containing either an OSM or an individual born to an OSM, whether or not they were members of the original sample. In practice, this means that split-off OSMs, such as children leaving the parental home or an ex-spouse who leaves the original household after a divorce, are followed in all subsequent waves. Similarly, these split-off persons are also followed when they reunite with members of their former households, as when a couple separates but then reunites or a child returns to the parental home after a "false" start in an independent household. However, TSMs are only followed in the subsequent wave on the condition that they continue to live with an original sample member.

The main advantage of household panel surveys is the rich set of variables that they provide. For example, the BHPS, a nationally representative sample of more than 5,000 households and 10,000 individual interviews, was originally designed as a research resource to further understanding of social and economic change at the individual and household levels in Britain. In addition, as the duration of the BHPS lengthens, new analyses become feasible. After more than 15 years of data collection, and with both parents and children as sample members in their own right, it becomes possible to conduct analyses of intergenerational influences and intergenerational mobility.

Household panels typically use the same sort of instruments to collect this rich set of variables about each household and its members. Questions about the household itself (e.g., type of dwelling, housing costs, enumeration of household members, and the relationships between the household members) are answered by the household head or some other designated adult. For some household panel surveys, this designated adult also responds to questions about the individual household members; however, in most household panel surveys each adult member of each household responds to an individual questionnaire that asks about personal characteristics and behavior. Typical components of this individual questionnaire are personal income, employment, health, time use, and attitudes on various subjects.

Rotating Panel Surveys

A repeated panel survey is made up of a series of individual panel surveys. When there is overlap in the time periods covered by the individual panels, and individual panel members are rotated into and out of the panel over a relatively short time period, we speak of a *rotating panel survey*. An initial sample of respondents is selected and interviewed for a predetermined time, from a few months to several years, but at intervals shorter than for most household panels. During the life of this first panel, a new panel is selected, followed, and interviewed in the same way as in the first panel. Third and subsequent panels are constructed similarly. Each individual panel has a pre-determined fixed life, although the overall rotating panel survey usually has an indefinite life.

Rotating panels are used to provide a series of cross-sectional estimates (e.g., unemployment rates and changes in those rates), but they also have a focus on short-term longitudinal measures (e.g., durations of spells of unemployment). For example, the Survey of Labour and Income Dynamics (SLID) provides national data on a whole range of transitions, durations, and repeated occurrences of individuals' financial and work situations. The SLID is a Canadian rotating panel that started in 1993, consisting of a succession of overlapping panels each with a duration of 6 years and with each new panel introduced 3 years after the introduction of the previous one. Each panel of the SLID consists of roughly 15,000 households and about 30,000 adults, and respondents are interviewed annually.

The rules for following respondents in rotating panels are similar to those in household panels. However, by restricting the duration of each panel to a shorter period, problems of attrition are reduced and representativeness is more easily maintained. In addition, using the combined data from the constituent panels with overlapping measurement periods, rotating panel surveys as a whole can provide better cross-sectional information at each point in time. Another advantage of rotating panel surveys is that the shorter interval between the interviews can reduce recall error. The longer the time between waves, and thus, the longer the reference period, the more recall errors that occur. For example, the U.S. Survey of Income and Program Participation (SIPP), with individual national panel with a sample size ranging from approximately 14,000 to 36,700 interviewed households and a duration that ranges from two and a half years to four years, interviews panel members every 4 months and uses a recall period of 4 months to collect data about the source and amount of income, labor force information, program participation, and eligibility data.

Femke De Keulenaer

See also Attrition; Longitudinal Studies; Panel; Panel Data Analysis; Rotating Panel Design; Wave

Further Readings

Kalton, G., & Citro, C. F. (1993). Panel surveys: Adding the fourth dimension. *Survey Methodology*, *19*(2), 205–215.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (Eds.). (1989). *Panel surveys.* New York: Wiley.

Paper-and-Pencil Interviewing (PAPI)

Prior to the 1980s, essentially all survey data collection that was done by an interviewer was done via paper-and-pencil interviewing, which came to be known as PAPI. Following the microcomputer revolution of the early 1980s, computer-assisted interviewing (CAI)—for example, computer-assisted personal interviewing (CAPI), computer-assisted self-interviewing (CASI), and computer-assisted telephone interviewing (CATI)—had become commonplace by the 1990s, essentially eliminating most uses of PAPI, with some exceptions. PAPI still is used in instances where data are being gathered from a relatively small sample, with a noncomplex questionnaire, on an accelerated start-up time basis, and/or the time and effort it would take to program (and test) the instrument into a computer-assisted version simply is not justified. PAPI also serves as a backup for those times when computer systems go down and interviewers would be left without work if there were not a paper version of the questionnaire to fall back to on a temporary basis. (Of note, mail surveys typically use paper-and-pencil questionnaires, but since they are not interviewer-administered surveys, mail questionnaires and that mode of data collection are not discussed here.)

PAPI is markedly inferior to CAI in many ways. The most important of these are (a) how sample processing is done with PAPI and (b) the limits of the complexity of the questionnaires that can be implemented via PAPI. Processing sample cases in PAPI traditionally was done manually. This required a supervisory person or staff to hand-sort "call sheets" or "control sheets" that were printed on paper, on which the interviewers filled out information each time an attempt was made to complete a questionnaire with a sampled case (e.g., at a telephone number or household address). This manual approach put practical limits on the complexity of the sample management system that could be used to sort and reprocess the active sample. It also relied entirely on the behavior and memory of the sample coordinator, which of course was fallible.

Questionnaires in PAPI cannot practically deploy complex randomization schemes that are easily programmed into and controlled by CAI. Although randomization can be built into PAPI, it typically requires that multiple versions of the questionnaire be created, printed, and randomly assigned to sampled cases. And, while randomized "starts" to question sequences can also be implemented in PAPI, interviewer error in implementing that type of randomization accurately is much more prevalent in PAPI. True randomization of the order of items within a question sequence is a nightmare to implement—if not outright impossible—accurately in PAPI when there are more than two items to randomize. The use of questions that use "fills" from answers previously given by the respondent (e.g., *Earlier you said that you had gone to the hospital X times the past 3 months . . .*) is also much more difficult to implement accurately in PAPI, whereas there are essentially no limits to its use in CAI. PAPI also has no assured way to control an interviewer from entering an "out-of-range" value to a particular question, whereas in CAI valid value ranges are programmed into each question asked.

The legibility of answers recorded by interviewers to open-ended questions in PAPI always is more problematic than what is captured via CAI. All in all, there is a great deal more potential for certain types of interview-related error in data collection in PAPI than is the case with CAI.

Data processing with PAPI is more time consuming and error prone than with CAI, given the answers that are recorded on paper by interviewers must be transformed into some computer-readable format via data keypunching. Data archiving also is much more problematic with PAPI, as even a small-sized survey organization must store boxes and boxes of completed questionnaires for some period of time after the completion of a survey project. However, although computer files can become corrupted and important past data can be lost, PAPI questionnaires are rarely destroyed, unless a fire or water damage occurs, until they are purposely thrown away.

There are some important benefits that occur from PAPI and do not routinely occur with CAI that have been forgotten or not realized by many researchers, especially those who started their careers after 1990. The most important of these is the level of attention that the average interviewer pays to the questions being asked and the answers being given in PAPI. In PAPI, there is a greater cognitive burden on the interviewer to negotiate the questionnaire, and a successful PAPI interviewer quickly learns that she or he must pay close attention to what she or he is asking the respondent and to what the respondent is replying. Often, this causes good PAPI interviewers to be more alert to problems and inconsistencies a respondent is creating than is typical of successful CAI interviewers. The PAPI interviewers then can try to probe to clarify or otherwise remedy these issues. In CAI, too many "successful" interviewers appear to take a mindless approach to interviewing and simply allow the computer lead them through the interview without really paying attention to the substance of what the respondent is saying. PAPI does not guarantee that this benefit will result, nor does CAI guarantee it will not result, but with PAPI, experience shows that interviewers are more likely to be intellectually engaged in the interviewing task than with CAI.

A second "benefit" of PAPI over CAI is that the same questionnaire, all other factors being equal, is completed more quickly if the interviewer uses a paper version of the questionnaire. Some have estimated the time difference to be 10%–20% longer with CAI than with PAPI. The reason underlying this phenomenon is that PAPI interviewers have much more active control over the pace of going from question to question than do interviewers using CAI. In PAPI, the "next" question generally is on the same page right below the current question, and the interviewer does not have to wait for the computer to display the next question before going on to it. In fact, in PAPI, interviewers often start reading the next question as they are recording the answer to the previous question. This does not occur as readily in CAI, since the computer software does not display the next question until after the answer to the current question has been entered by the interviewer.

Paul J. Lavrakas

See also Call Sheet; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Control Sheet; Interviewer-Related Error; Random Order; Random Start; Sample Management

Paradata

Paradata, also termed *process data* (but not to be confused with metadata), contain information about the primary data collection process (e.g., survey duration, interim status of a case, navigational errors in a survey questionnaire). Paradata can provide a means of additional control over or understanding of the quality of the primary data (the responses to the survey questions).

Collecting Paradata

Since paradata are defined simply as data describing the primary data collection process, paradata can be collected in every survey mode. However, the amount, type, and level of detail of the captured paradata will vary depending on whether the data have to be manually recorded or whether they are automatically logged by computer software. A crude distinction can also be made between paradata describing the data collection process as a whole (calls, follow-up procedures, etc.) and more specific paradata referring to how a survey questionnaire was filled in.

Case management software (such as that used in centralized computer-assisted telephone interviewing [CATI] facilities) can record a wide array of paradata about the survey process as a whole. The software is capable of logging the time, duration, and outcome of each call to a sample unit, although some of this information may need to be supplied by the interviewer (e.g., call outcome). In noncomputerized settings, paper call sheets filled in by the interviewers can serve to collect paradata.

The most efficient way of collecting paradata on how a survey questionnaire is filled in is to use a computerized survey questionnaire with software that logs meaningful actions such as ticking response options, navigating through the questionnaire, and so on. This could be viewed as an automatic behavior-coding system.

Whether and which paradata can be collected depends on the software used to create the survey questionnaire. Many CAI software packages allow the recording of paradata. For Web surveys, JavaScript code has been developed to collect detailed paradata similar to keystroke data generated by CAI software. (Apart from these data, Web server logs can also be used to collect less-detailed paradata.)

Uses of Paradata

Paradata can assist survey questionnaire pretests. For instance, data on how long it took to answer survey questions could be of importance in this phase. Long response latencies could indicate problems with particular questions. Paradata from keystroke files can reveal where errors were made, which may indicate poor interface design.

Paradata can also be collected during the actual field work. Recently, researchers have used paradata to adapt the survey design while the field work is still ongoing in order to improve survey cost efficiency and to achieve more precise, less biased estimates (these are so-called responsive design surveys).

In interviewer-administered surveys, paradata can be used to evaluate interviewer behavior. Time data can help identify interviewers who administered all or parts of the questionnaire too quickly. As in pretests, keystroke data can reveal where errors are being made. If these analyses are conducted during the field work, corrective measures can still be implemented in this phase.

When conducting experiments in a survey (e.g., within the questionnaire), researchers can use paradata as an additional source of information about the effects of the experimental treatment. Apart from a test of the effect of the treatment (e.g., survey instructions) on the dependent variable (e.g., omission errors in a self-administered questionnaire), paradata allow the researcher to see the effects of the treatment on the response behavior itself. This may provide additional insight into the reason why a specific treatment is effective or not.

Data Preparation

If the researcher has clear a priori assumptions about which information is important, paradata can be stored in conventional matrix form data sets. This, for instance, is the case when call sheets are used or when simple survey question durations are recorded. The variables are pre-defined, and the interviewer or computer software is instructed to compute and record their values. These variables can then be used in conventional analyses.

If some variables cannot be pre-defined, the data can be collected in a relatively unstructured way. Keystroke data are of this type. Data are collected by adding each action to a data string. Since not every interviewer or respondent will perform the same number of actions, or in the same sequence, these data strings will be of different lengths and their structure will vary from one observation to the next. In addition, different parts of the data strings could be important depending on the focus of the analysis. SAS macros or other software capable of recognizing string patterns can be used to extract the useful information from the strings before the actual analysis.

Dirk Heerwegh

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Response Latency; SAS; Web Survey

Further Readings

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research: A method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, *55*, 331–346.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*, 439–457.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client side paradata from a Web survey. *Social Science Computer Review*, *21*, 360–373.

Jeavons, A. (1999). Ethology and the Web: Observing respondent behaviour in Web surveys. *Marketing & Research Today*, *28*, 69–76.

Parameter

A parameter is a numerical quantity or attribute of a population that is estimated using data collected from the population. Parameters are to populations as statistics are to samples. For example, in survey research, the true proportion of voters who vote for a presidential candidate in the next national election may be of interest. Such a parameter may be estimated using a sample proportion computed from data gathered via a probability sample of registered voters. Or, the actual annual average household "out-of-pocket" medical expenses for a given year (parameter) could be estimated from data provided by the Medical Expenditures Survey. Or, the modal race of students within a particular school is an example of an attribute parameter that could be estimated using data acquired via a cluster sample of classrooms or students from the particular school. An important parameter in the realm of survey nonresponse is the likelihood of the response. The binary event "respond or not respond" can be modeled as a Bernoulli random variable with unknown parameter, p, which can vary by sampling unit as a function of demographic, socioeconomic, or other variables.

Parameters may also refer to specific aspects of a sampling distribution of a test statistic or reference distribution. For example, when estimating the mean weight loss for subscribers of a particular weight loss plan (the parameter), the most straightforward point estimate is the sample mean. The corresponding confidence interval is then computed with respect to a reference distribution—usually approximated by a normal distribution with a location and a scale parameter. The location parameter—or mean of the sampling distribution of the sample mean—has the same value as the population mean (parameter); the scale parameter, or standard deviation, is equal to the population standard deviation divided by the square root of the sample size. In general, the statistical parameters for the approximate sampling distribution of a statistic end up being equal to a function of the actual population parameters themselves.

While parameters are generally of direct interest in both estimation and inference, they can also serve as "auxiliary" inputs for statistical techniques to improve estimates of target or primary parameters. For example, in the most current random-digit dial (RDD) survey practice, landline RDD samples of households may be augmented with a screened sample of cell phone numbers identified as "cell phone only." A reliable estimate of the distribution of "type of phone ownership" (a nontarget parameter) is then used to adjust the initial survey weights via raking techniques to provide overall unbiased estimates of these population totals. The true number of patients diagnosed with a particular stage of cancer as well as the actual number of patients of each gender within a state are auxiliary parameters or universe counts that can be used to adjust the base survey weights so that representation by stage and sex may be achieved for a state registry-based survey that aims to estimate the percentiles of the distribution of quality-of-life scores for patients diagnosed with lung or skin cancer. These auxiliary parameters can also be used in calibration estimators to adjust for survey nonresponse.

Parameters can be univariate, bivariate, or multivariate quantities that, in turn, can be estimated appropriately by univariate, bivariate, or multivariate statistics. The regression parameters for the impact of income and education on average number of hours spent watching television give an example of several parameters that are estimated simultaneously from sample data via finite population regression. Multivariate collections of parameters can also be more complex, to include variances and covariances of a collection of measured variables contained in a survey questionnaire, along with path coefficients or factor loadings of latent variables in the context of factor or latent class analysis.

Trent D. Buskirk

See also Confidence Interval; Nonresponse; Point Estimate; Population; Raking; Sample; Statistic

Further Readings

Yates, D., Moore, D., & McCabe, G. (1999). *The practice of statistics*. New York: W. H. Freeman.

Partial Completion

The partial completion survey disposition is used in all types of surveys, regardless of mode. In a telephone or in-person interview, a partial completion results when the respondent provides answers for some of the questions on the survey questionnaire that were asked by the interviewer but is unable or unwilling to allow the interviewer to administer all of the questions in the interview (item nonresponse). Partial completions in telephone or in-person surveys can occur when an appointment or other commitment prevents the respondent from completing the interview or when the respondent begins the interview but then refuses to complete the entire interview process (called a "breakoff"). In a mail survey, a partial completion results when the respondent receives a paper-and-pencil survey questionnaire, answers only some of the questions on the questionnaire, and returns the questionnaire to the researcher. In an Internet survey, a partial completion occurs when the respondent logs into the survey, enters answers for some of the questions in the questionnaire, and submits the questionnaire electronically to the researcher. If a partial is not a hostile breakoff, most survey firms attempt to recontact the respondent who completed the partial interview (by telephone, mail, or Internet, depending on the survey mode) in order to attempt to get a completed interview.

In practice, the difference between completed interviews, partial completions, and breakoffs is that completed interviews contain the smallest number of item nonresponses, while breakoffs contain the largest number of item nonresponses. Most survey organizations have developed rules that explicitly define the difference among breakoffs, partial completions, and completed interviews. Common rules used by survey organizations to determine whether an interview with item nonresponse can be considered a completed interview include (a) the proportion of all applicable questions answered, and (b) the proportion of critically important or essential questions administered. For example, cases in which a respondent has answered fewer than 50% of the applicable questions might be defined as breakoffs; cases in which the respondent has answered between 50% and 90% of the applicable questions might be defined as partial completions; and cases in which the respondent has answered more than 90% of applicable questions might be considered completed interviews.

Matthew Courser

See also Breakoff; Completed Interview; Final Dispositions; Item Nonresponse; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Percentage Frequency Distribution

A percentage frequency distribution is a display of data that specifies the percentage of observations that exist for each data point or grouping of data points. It is a particularly useful method of expressing the relative frequency of survey responses and other data. Many times, percentage frequency distributions are displayed as tables or as bar graphs or pie charts.

The process of creating a percentage frequency distribution involves first identifying the total number of observations to be represented; then counting the total number of observations within each data point or grouping of data points; and then dividing the number of observations within each data point or grouping of data points by the total number of observations. The sum of all the percentages corresponding to each data point or grouping of data points should be 100%. The final step of creating a percentage frequency distribution involves displaying the data.

For example, as part of a study examining the relationship between number of trips to a physician and socioeconomic status, one might survey 200 individuals about the number of trips each made to a physician over the past 12 months. The survey might ask each individual to choose from the following responses: "0 times during the past year," "1–3 times during the past year," "4–6 times during the past year," "7–9 times during the past year," and "10 or more times during the past year."

If 10 respondents were to select the first response, 80 were to select the second, 50 were to select the third, 40 were to select the fourth, and 20 were to select the fifth, then the percentage frequency distribution would be calculated by dividing the number of responses for each choice by the total number of responses, or 200. The percentage frequency of each would be 5%, 40%, 25%, 20%, and 10%, respectively. The percentage frequency distribution is shown in table form (Table 1) and in bar graph form (Figure 1).

Alternatively, one could aggregate—or group—data points. For instance, in the previous example, a percentage frequency distribution could group the number of trips to the doctor into three distinct categories, such as "0 times," "1 to 6 times," and "7 or more times." When grouping the responses, the total percentage of each category of response is merely the sum of the percentages for each response. When grouped in this manner, the frequency percentage of "0 times" would remain unchanged at 5%, the total for "1 to 6 times" would be 65% (the sum of 40% and 25%), and the total for "7 or more times" would be 30% (the sum of 20% and 10%).

Most statistical software programs can easily generate percentage frequency distributions and provide visual representations in table or graph form.

Joel K. Shapiro

See also Frequency Distribution; Relative Frequency; Unit of Observation

Further Readings

Kvanli, A. H., Guynes, C. S., & Pavur, R. J. (1986). *Introduction to business statistics*. St. Paul, MN: West.

Table 1 Percentage frequency distribution

Response	*Percentage (%) of Respondents Selecting Response*
0 times during the past year	5
1–3 times during the past year	40
4–6 times during the past year	25
7–9 times during the past year	20
10 or more times during the past year	10

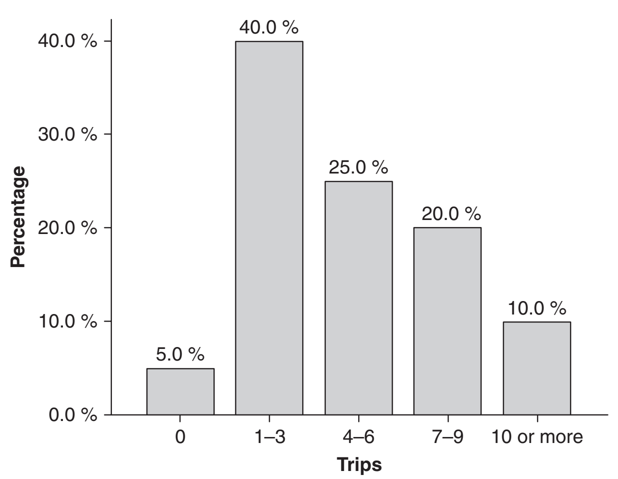

Figure 1 Percentage frequency distribution

Percentile

A percentile is a statistic that gives the relative standing of a numerical data point when compared to all other data points in a distribution. In the example $P_{.84} = 66$, $P_{.84}$ is called the *percentile rank* and the data point of 66 is called the *percentile point*. The .84 in the percentile rank of $P_{.84}$ is a proportion that tells us the relative standing of the percentile point of 66 compared to all other data points in the distribution being examined. Reporting percentiles can be a useful way to present data in that it allows an audience to quickly determine the relative standing of a particular data point.

By itself, a raw score or data point says little about its relative position within a data set. Percentiles provide a number expressing a data point's relative position within a data set. At a glance, the percentile shows the reader whether a particular numerical data point is high, medium, or low in relation to the rest of the data set. Salaries, IQ scores, standardized test scores such as the SAT, GRE, body mass index (BMI), height, and weight are all frequently expressed as percentiles.

Some percentile values commonly used in reporting are the median, $P_{.50}$, below which 50% of the cases fall; the lower quartile, $P_{.25}$, below which 25% of the cases fall; and the upper quartile, $P_{.75}$, below which 75% of the cases fall. The area between the lower quartile and the middle quartile is called the "interquartile range," which contains the middle 50% of values in a distribution.

There are two basic definitions of the proportion expressed in the percentile rank. One definition used in some introductory statistics textbooks calculates the percentile rank as the proportion of cases falling *below* the percentile point. Using this definition, the maximum obtainable percentile must be less than 1.0, because there is no number in a data set that falls below itself. The second definition of percentile rank is the proportion of cases *at or below* the percentile point. Using this second definition, the 100th percentile is the maximum obtainable percentile, because 100% of the data falls at or below the largest number in a data set. The definition of *percentile* is dependent on the formula used to calculate the percentile rank.

Using our example $P_{.84} = 66$, the first definition of percentile rank calculates the percentile rank of .84 to mean 84% of the cases in the distribution fall *below* the percentile point of 66. A relatively simple way to calculate percentiles using this definition can be obtained with the formula $p(N)$ where p is the desired percentile rank and N is the number of cases in the distribution. This calculation gives the position within the distribution where the percentile point is located once the data points in the distribution are ordered from lowest to highest. If $p(N)$ results in a fractional number, round up to the next highest number for the percentile point position within the distribution. Once the position within the data set is determined, count up from the bottom of the distribution to the number obtained from the calculation $p(N)$. The mean of that number in the data set and the number value in the next highest position in the distribution is the percentile point corresponding to the percentile rank.

The calculation given is by no means the only way to calculate percentiles; however, it is one of the simplest. Statistical software programs such as Statistical Package for the Social Sciences (SPSS) and SAS and spreadsheet programs such as Microsoft Excel allow the user to calculate percentiles quickly. SPSS and SAS allow the user to choose from a variety of different formulas that will calculate the percentile values in slightly different ways, yielding slightly different results, depending on the user's needs.

Dennis Dew

See also Mean; Median; SAS; Stata; Statistic; Statistical Package for the Social Sciences (SPSS)

Further Readings

McCall, R. B. (2001). *Fundamental statistics for behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Perception Question

Perception is the subjective process of acquiring, interpreting, and organizing sensory information. Survey questions that assess perception, as opposed to those assessing factual knowledge, are aimed at identifying the processes that (a) underlie how individuals acquire, interpret, organize, and, generally make sense of (i.e., form beliefs about) the environment in which they live; and (b) help measure the extent to which such perceptions affect individual behaviors and attitudes as a function of an individual's past experiences, biological makeup, expectations, goals, and/or culture.

Perception questions differ from other types of survey questions—behavioral, knowledge, attitudinal, or demographic—in that questions that measure perception ask respondents to provide information on how they perceive such matters as the effectiveness of programs, their health status, or the makeup of their community, among other specific measures assessing biological, physiological, and psychological processes.

Broadly, research on perception is driven by many different kinds of questions that assess how individual senses and perceptions operate; how and why individuals are susceptible to perceptions or misperceptions; which structures in the brain support perception; and how individual perceptions acquire meaning. Research on the psychology of perception suggests that the actions of individuals are influenced by their perceptions of the opinions, values, and expectations of others, including those individuals identified as important by the respondent. This is of particular import to survey methodologists, because an individual's perceptions may influence her or his survey responses and, moreover, may be inaccurate. When this inaccuracy is systematic (biasing) rather than random, such inaccuracy has consequences for interpreting the survey data collected. Theory development on perception indicates that the accuracy of reported perceptions is related to communication mode, coordination efforts, and the salience of the percept. Other research finds that perceptions may be distorted by intimate relationships, attraction, personality traits, or indirect cues or interviewer effects, among other influences. Such influences may induce social desirability bias, where perception questions are improperly drafted or fail to account for question wording and ordering effects within the instrument and administration mode.

Including perception measures in a survey instrument enables researchers to investigate both qualitative and quantitative empirical hypotheses by incorporating open-ended and closed-ended measures that assess the way the respondent acquires, interprets, and organizes information, questions about the relationships among the respondent's perceptions, and the meaning of reported perceptions. To this end, introspection, experimental psychology, and neuroscience research are used to study perception in fields ranging from cognitive to computer science; each use different questions to measure perception. The method of introspection asks respondents to examine and report their conscious thoughts, reasoning, or sensations, such as the questioning used in cognitive interviewing. Here, respondents may be asked how their perceptions compare with the perceptions of other people.

The subjective nature of perception, however, presents a reliability problem. Because the survey interviewer or researcher cannot often easily or reliably identify whether a respondent is being truthful or accurate in her or his reports of a subjective experience, it is not possible to tell whether a particular word used to report an experience is being used to refer to the same kind of experience reported by another respondent. Thus, the reliability of introspection and the data collected using this method should be scrutinized carefully. Methods in experimental psychology include questions that prompt the participant to report, for example, object recognition, motion detection, visual illusions, and the like. In some cases, this report follows the participant performing a perceptual task. Neuroscience research uses perception questions to study attention and memory systems to identify how individuals store, organize, and retrieve perceptions as a way to understand information processing generally.

Finally, it is important, given the subjective nature of perception questions, to minimize the error in survey statistics by choosing wisely those design and estimation methods that are likely to reduce error.

Traci Lynne Nelson

See also Behavioral Question; Cognitive Aspects of Survey Methodology (CASM); Cognitive Interviewing; Construct; Construct Validity; Context Effect; Experimental Design; Gestalt Psychology; Interviewer Effects; Knowledge Question; Measurement Error; Question Order Effects; Respondent-Related Error; Saliency; Social Desirability

Further Readings

Blake, R., & Sekuler, R. (2005). *Perception* (5th ed.). Columbus, OH: McGraw-Hill.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (1989). *Survey methodology*. New York: Wiley.

PERTURBATION METHODS

Perturbation methods are procedures that are applied to data sets in order to protect the confidentiality of survey respondents. The goal of *statistical disclosure control* (SDC) is to provide accurate and useful data—especially public use data files—while also protecting confidentiality. Various methods have been suggested, and these may be classified two ways: (1) methods that do not alter the original data but reduce the amount of data released; and (2) methods that alter individual values while maintaining the reported level of detail. The first set of methods may be described as *data coarsening;* the second set of methods may be described as *statistical perturbation methods*.

Perturbation methods have the advantage of maintaining more of the actual data collected by survey respondents than data coarsening. Variables selected for perturbation may be those containing sensitive information about a respondent (such as income) or those that may potentially identify a respondent (such as race). These methods can be used for data released at the microdata level (individual respondent records) or at the tabular level (in the form of frequency tables). Depending on the data, their values, and method of data release, researchers may select one perturbation method over another, use multiple perturbation techniques, or use these techniques in addition to data coarsening.

Examples of perturbation methods are described below, with a focus primarily on perturbation of microdata. This is not an exhaustive list, as new methods are continually being developed.

Data swapping. In this method, selected records are paired with other records in the file based on a predetermined set of characteristics. Data values from some identifying or sensitive variables are then swapped between the two records. The sampling rate is designed to protect the confidentiality of the data without affecting the usability of the data set. This method introduces uncertainty to an intruder as to which reported values were provided by a particular respondent.

Rank swapping, a method similar to data swapping. With rank swapping, pairs are created that do not exactly match on the selected characteristics but are close in terms of the ranks of the characteristics.

Adding random noise. This method is a way of masking sensitive items by adding or multiplying by random numbers. The random numbers are selected from a pre-specified distribution with a mean of 0 and a selected standard deviation, so that the value is altered as little as possible but enough to prevent reidentification.

Replacing values with imputed data. With this method, specified sensitive values on a randomly selected set of records are replaced with imputed values from other, similar records. This approach will introduce some uncertainty as to whether the sensitive items on a record were actually reported by a particular respondent.

Data synthesis. Values are replaced with those predicted from models developed to generate multiple imputations that allow for valid statistical inference. All values for all records may be replaced (full synthesis), or a subset of variables on a subset of records (partial synthesis).

Blurring. In this method, small groups of records are formed based on the proximity of their values of a sensitive variable or other variables related to the sensitive variable. Aggregate (usually average) values are calculated from the individual responses for the sensitive item in that group. The aggregate value may be used in place of one (for example, the middle) or all individual responses for the group on the released data file.

Microaggregation. This is similar to blurring. However, the records are grouped so they are similar in terms of all sensitive variables of interest. This same grouping is used to "blur" all the sensitive variables.

Supersampling and subsampling. Records from the original file are sampled with replacement and added to it. The result is a file that is larger than the original microdata file. The larger file is then subsampled to produce the final microdata file. This method reduces the appearance of actual sample uniques (since they may not be sampled for the final file) and creates others (since a value and its duplicate may not both be selected).

Post-randomization method (PRAM). For each record on the data file, the values of one or more categorical variables are changed to already existing values on the file. These changes are made independently of other records using a pre-determined probability distribution. The level of protection depends on the probability matrix, the values, and their frequencies in the original data.

Data shuffling. This method is similar to data swapping. However, rather than have data values exchanged between records, a value for a sensitive variable on record a is replaced with the value from record b. The value from record b is then replaced with the value from record c, and so on, based on the conditional distribution of the sensitive variable.

Rounding. This is a perturbation method appropriate for tabular data. With random rounding, the decision to round a value up or down (to a pre-determined base) is made at random. Controlled rounding is similar, but the adjustments are such that the original marginal totals of the table are preserved.

Sylvia Dohrmann

See also Confidentiality; Data Swapping; Privacy; Suppression

Further Readings

Elliot, M., Hundepool, A., Nordholt, E. S., Tambay, J.-L., & Wende, T. (2006). The new glossary on statistical disclosure control. *The Survey Statistician*, *54*, 26–36.

Federal Committee on Statistical Methodology, Subcommittee on Disclosure Limitation Methodology. (1994). *Report on statistical disclosure limitation methodology*. Statistical Policy Working Paper 22. Washington, DC: Office of Management and Budget, Executive Office of the President.

PEW RESEARCH CENTER

The Pew Research Center is a nonpartisan research center based in Washington, D.C. There are seven separate projects within the Pew Research Center, most of which employ sample surveys as a primary research tool. In addition to relying on surveys for much of its research, the Pew Research Center also conducts research on survey methodology.

The oldest of the projects is the Pew Research Center for the People & the Press, which was founded in 1987 as the Times Mirror Center for the People & the Press. The Times Mirror Center was originally created by the Times Mirror Company, a media corporation that owned *The Los Angeles Times* and other major newspapers and communications properties. Its mission was to conduct in-depth research that would illuminate the connections among the public policy world, the press, and the public, and to disseminate the research widely and without cost. In 1995, the Pew Charitable Trusts became the primary source of funding for the center, which was renamed the Pew Research Center for the People & the Press. In 2004, several other Pew-funded research projects were combined under the umbrella organization of the Pew Research Center. The Pew Charitable Trusts is a Philadelphia-based public charity.

The president of the Pew Research Center is pollster Andrew Kohut, who was president of the Gallup Organization from 1979 to 1989 and who served as the research director for the Times Mirror Center when it was founded. He subsequently became its director in 1993. When funding from Times Mirror ended, Kohut obtained funding from Pew to continue the center's operations. In subsequent years, he oversaw the creation of the Pew Internet and American Life Project, which extended the People & the Press's previous work on the Internet. He also forged research partnerships with the Pew Forum on Religion & Public Life and the Project for Excellence in Journalism, Pew-funded projects that are now part of the Pew Research Center.

The Pew Research Center for the People & the Press conducts monthly polling in the United States on policy issues, public interest in the news, and political parties and elections. Among its regular but less frequent projects are (a) a biennial survey on news consumption; (b) a survey of foreign policy attitudes among the public and several groups of foreign policy elites; (c) a survey of U.S. journalists, conducted with the Project for Excellence in Journalism, regarding issues facing the news industry and journalism; (d) an annual survey on religion and politics conducted with the Pew Forum on Religion & Public Life; (e) surveys on the political values of the public; and (f) methodological research focused on issues facing the survey research community. The center also conducts extensive polling during national elections, typically including a broad "scene-setter" poll during the summer prior to the election, several polls during the fall focused on the issues and candidate images, a predictive survey conducted the weekend before the election, and a post-election survey to gauge the public's reaction to the outcome of the election and expectations for the future. In conjunction with the Pew Internet & American Life Project, the People & the Press Center also conducts a post-election survey on the public's use of the Internet to follow and participate in the election campaigns.

The Pew Internet & American Life Project, founded in 1999, uses surveys to study the social and political impact of the Internet in the United States. Most of its work consists of random-digit dialed (RDD) telephone surveys of the general public and special populations, but it also uses Internet surveys and qualitative methods in some of its research. Its director is Lee Rainie, a former journalist who served as managing editor of *U.S. News & World Report.* The project conducts regular tracking surveys to monitor trends in Internet use for a range of activities, including email use, broadband adoption, search engine use, blog creation and readership, and use of the Internet in such areas as health care, hobbies, the arts, social interaction, education, shopping, decisions about major purchases, and political activity.

The Pew Hispanic Center makes extensive use of survey research to study the Hispanic population of the United States and its impact on the country. The center conducts regular RDD telephone surveys of the Hispanic population, focusing on such topics as political engagement, employment, identity, education, and remittances. It also conducts extensive secondary analysis of U.S. Census surveys such as the decennial census, the Current Population Survey, and the American Community Survey. Its founding director is Roberto Suro, a former journalist who served as reporter for *The Washington Post, Time* magazine, and *The New York Times*.

In addition to partnering with the People & the Press Center on annual surveys of religion and politics in the United States, the Pew Forum on Religion & Public Life also had used surveys to study Pentecostalism in Asia, Latin America, and Africa as well as in the United States. It also has worked with the Pew Hispanic Center on a survey of Hispanics regarding religious beliefs and practices and is undertaking an effort to track the size and composition of religious groups in the United States. The Pew Forum's director is Luis Lugo, a former professor of political science and director of the religion program for the Pew Charitable Trusts.

The Global Attitudes Project, founded in 2002 by Andrew Kohut, conducts public opinion polling internationally on a range of topics including politics, religion, economics, social life, and foreign affairs. Among the subjects of its surveys has been the rise of anti-American sentiment, opinions about Islamic extremism, views of democracy among Muslim publics, and opinions about the impact of globalization.

Scott Keeter

Further Readings

Pew Research Center: http://pewresearch.org

Pilot Test

Pilot tests are "dress rehearsals" of full survey operations that are implemented to determine whether problems exist that need to be addressed prior to putting the production survey in the field. Traditional pilot tests are common and have been a part of the survey process since the 1940s. In recent years, by the time a pilot test is conducted, the questionnaire has frequently already undergone review (and revision) through expert review, focus groups, and/or cognitive interviews.

The terms *pretest* and *pilot test* are sometimes used interchangeably; however, in recent years *pretest* has taken on the meaning of testing within a survey laboratory, rather than in the field with the general population. Some organizations or survey researchers now refer to pilot tests as *field pretests*. Pilot testing is one of the most critical aspects of a successful survey operation resulting in good survey data. Going into the field for a full production survey without knowing whether the questionnaire and/or field interviewer procedures work is a recipe for disaster.

Objectives

In surveys, nonsampling measurement error can be caused by problems associated with interviewers, respondents, and/or the questionnaire. The objective of a pilot test is to identify potential problems and address them prior to the production survey to reduce the amount of nonsampling measurement error produced by the survey.

Procedures

The pilot test procedures should mirror the procedures that will be used in the production survey. For pilot tests, the sample size is typically 50–100 cases. In fact, *sample* is an inappropriate term to use, since a nonrandom convenience sample rather than a random sample is typically used. It is important to have a "sample" that is as similar in characteristics as the respondents in the production survey sample. But due to costs and staff efficiencies, pilot tests are frequently done in one to three locations in the country, in particular when data collection is face-to-face, with a relatively small interview staff. If a project involves surveying persons with unique characteristics, it is extremely important that persons with the targeted characteristics be included in the pilot test sample.

All of the procedures used in the production survey should be used in the pilot test. This includes modes of survey administration, respondent rules, interviewer staffing, and interviewer training.

It is not beneficial to include only the best or most experienced interviewers in the pilot test. Those interviewers often have enough experience that they know how to make a problematic question work, but their solutions sometimes lie outside of standardized interviewing practices and are therefore inconsistent from case to case. Using some inexperienced and/or low-caliber interviewers allows problematic situations to arise naturally and be evaluated. If the situation arises during the pilot test, it is likely it will be encountered in the production survey. It is better to find out about the problem during the pilot test, so that the problem can be addressed prior to production interviewing.

Implementation

When the pilot test is implemented, an estimate of average interview time for questionnaire completion can be obtained that has implications on the survey budget. If the interview time exceeds that which is allowed or budgeted for, then decisions will need to be made about reducing the number of the questions in the survey, reducing the sample size, or changing interview procedures. If there are deep or rare paths of questioning in the questionnaire, reducing the sample size may result in not having enough cases for meaningful analyses of the characteristics associated with the rare paths. The pilot test will also provide information about question comprehension, sensitivity, difficulty, and/or item nonresponse related to specific questions. The pilot test permits issues related to question sequencing and transitional lead-ins to surface. It is important to know whether a lead-in works to segue between questionnaire sections or to serve as a buffer just prior to sensitive questions. Information on case or unit refusals as well as item refusals or other nonresponse is produced through pilot tests. The pilot test provides information about interviewer difficulties related to survey administration and also improvements that may be needed for interviewer training.

Evaluation Methods

The most common form of evaluation of a pilot test is interviewer debriefings. Interviewer debriefings usually consist of the pilot test interviewers meeting together and along with the project manager; they go through the questionnaire question by question, identifying what problems they encountered. Feedback on interviewer training and interviewer procedures is also solicited, so that revisions can be made prior to the production survey. Relying solely on interviewers for feedback about the questionnaire is insufficient to gain an objective view, however. Some interviewers may have difficulty separating the problems the respondents encountered from issues about the questionnaire that the interviewers do not like. And some interviewers are more vocal than others, and the problems they perceive exist with the questionnaire may actually be less consequential than problems observed with different items by a less vocal interviewer.

Interviewer evaluation of pilot tests has expanded during the past 20 years to include standardized rating forms. With these forms, interviewers consider their pilot test experience and provide ratings as to whether, and the extent to which, the question was problematic. This allows all interviewers to weigh in equally on problem identification and provides empirical data for the researcher to analyze.

During the past 20 years, additional methods have been used to evaluate pilot tests. These include behavior coding, respondent debriefing, and vignettes. Behavior coding is used to evaluate interviewer and respondent interactions to determine if there are problems with specific questions. The information from behavior coding is even richer if coders make notes about the interviewer–respondent exchange whenever any problematic behaviors occur. The notes can assist the questionnaire designer when he or she is working to revise the question to improve it for the production survey. The notes often provide insights about the nature of the problem, and not just that a problem exists.

The respondent debriefing method has been used to obtain information about problematic concepts or questions to aid in identifying the problem source. A respondent debriefing consists of probing questions that are administered at the end of the interview to obtain additional information about earlier survey questions. Pilot tests provide the opportunity to probe for question comprehension from a larger and more representative sample of persons than laboratory testing due to greater sample selectivity with laboratory subjects than with pilot test respondents.

In some cases, the number of respondents in the pilot test is not large enough to adequately cover all the concepts and question paths in the questionnaire. With vignettes, short examples of specific situations related to critical concepts or questions in the survey instrument are presented to pilot test respondents. Respondents are asked how they would answer a particular question given the "story" in the vignette. Examining responses to vignettes helps to determine if a problem exists with specific questions or concepts of interest.

It is critical that the project schedule allow sufficient time for pilot test data evaluation so necessary revisions can be made and, if possible, retested to ensure that the revisions don't introduce a new set of problems. Time must be allowed for the benefit of the pilot test to be realized.

Implications

It is very important for survey organizations to build pilot testing into their timelines and budgets for each survey. Pilot testing is frequently the first activity to get cut when budget and time run tight, yet unknown problems of the questionnaire or interviewer procedures may lead to increased nonsampling measurement error and subsequently lower-quality survey data. Incorporating pilot tests into the routine procedures for survey development and planning is vital in order to identify weaknesses in the questionnaire and interviewer procedures prior to the production survey.

Jennifer M. Rothgeb

See also Behavior Coding; Convenience Sampling; Interviewer Debriefing; Measurement Error; Questionnaire Design; Questionnaire Length; Respondent Debriefing; Survey Costs

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Eds.). (2004). *Survey methodology.* Hoboken, NJ: Wiley.

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., et al. (Eds.). (1997). *Survey measurement and process quality.* New York: Wiley.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., et al. (Eds.). (2004). *Methods for testing and evaluating survey questionnaires.* Hoboken, NJ: Wiley.

Schwarz, N., & Sudman, S. (1996). *Answering questions.* San Francisco: Jossey-Bass.

Sudman, S., & Bradburn, N. (1982). *Asking questions.* San Francisco: Jossey-Bass.

Point Estimate

Point estimates are single numeric quantities (i.e., "points") that are computed from sample data for the purpose of providing some statistical approximation to population parameters of interest. For example, suppose surveys were being designed to estimate the following population quantities: (a) the proportion of teenagers within a school district who consumed at least one alcoholic beverage last year, (b) the mean number of candy bars consumed last week by county hospital nurses within a state, (c) the total number of text messages sent by cell phone customers of a particular cell phone provider within the last month, (d) the correlation between education and annual expenditures on magazine subscriptions within the past year for U.S. citizens. In every case, a single numeric quantity, or statistic, can be computed from collected sample data to estimate the population parameters of interest. In contrast to point estimates, *interval estimates* are computed using point estimates to provide an estimated range of values for the parameter.

Point estimates generally have a form that is consistent with the population parameter they are intending to estimate; for example, a sample mean is used to estimate a population mean; a sample proportion is used to estimate a population proportion; a sample correlation coefficient is used to estimate the population correlation coefficient. Within the context of survey research, point estimates can also be computed with or without survey weights. Moreover, point estimates are subject to sampling variability in that the values of the point estimates for a given parameter may vary from different samples of the same size selected from the same population.

For example, consider a sample of 10 students selected from a school district using a multi-stage probability sampling design to estimate the mean number of days absent from school during the most recent semester. These data are provided in Table 1.

One unweighted point estimate for the population parameter is given simply by the sample mean computed by dividing the sum of all 10 data points by 10. From the table, the sum of all the data points (sum of second column) is 47, so the unweighted point estimate given by the sample mean is 4.7 days absent for the semester. Because of the survey design, a weighted point estimate could also be computed using the Horvitz-Thompson estimator, which divides the weighted sum of all 10 data points by the sum of

Table 1 Student absentee data from a hypothetical sample of 10 students

Student	*Number of Days Absent Last Semester*	*Survey Weights*	*Weighted Values*
1	3	40	120
2	5	40	200
3	1	40	40
4	6	18	108
5	9	18	162
6	7	18	126
7	8	18	144
8	4	18	72
9	2	15	30
10	2	15	30
Column Sum	47	240	1032

the weights. From the table, the numerator of the weighted point estimate is 1,032 (column 4 sum) and the denominator is 240 (column 3 sum), so the weighted point estimate is given by $1032/240 = 4.3$ days absent for the semester.

Trent D. Buskirk

See also Interval Estimate; Population Parameter; Weighting

Further Readings

Feld, A. (2005). *Discovering statistics using SPSS.* Thousand Oaks, CA: Sage.
Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Brooks/Cole.

Political Knowledge

Political knowledge has long been considered an integral part of public opinion, as well as a topic of survey research interest in its own right for many decades. While many definitions of political knowledge exist, most reflect an understanding of it as factual information about politics and government that individuals retain in their memory.

Political knowledge is similar to, but somewhat narrower than, "political awareness," "political sophistication," or "political expertise." Political knowledge is an important source of the considerations people draw upon when asked to respond to an attitude or opinion question and is thus an integral aspect of many theories of public opinion formation and change. It also is a key concept in democratic theory, with scholars and philosophers since the time of Plato debating whether the public knows, or could know, enough to play an appropriate role in the governance of a society.

Political knowledge is typically measured in surveys with questions asking respondents to recall specific facts or to recognize names or events. There also are questions asking respondents to rate their own level of knowledge, either in general or on specific subjects. Related to these are questions that ask whether the respondent has heard or read anything about an event, person, or issue. Another type of measure is an interviewer rating of the respondent's level of knowledge. In the absence of these kinds of measures, researchers sometimes use other surrogates, such as level of formal education. While such surrogates may be correlated with political knowledge, they have significant limitations.

Knowledge questions began appearing on national surveys soon after the invention of the modern sample survey. One of the oldest knowledge questions appearing in the Roper Center's comprehensive database of survey questions is a January 1937 item that asked a national Gallup poll sample, *In your estimation, how much is the national debt today*? Soon thereafter, Gallup was regularly including knowledge questions on its national polls. Other survey organizations, including some located in government agencies, began asking knowledge questions in the late 1930s.

Despite the long interest by survey researchers in the concept, there was little sustained scholarly attention to the measurement of knowledge until the 1980s. Vincent Price's 1999 review of the topic for a major handbook on the measurement of political attitudes notes that the 1968 edition of the handbook identified only four developed scales, with limited evidence of reliability and validity. By 1999, there was an increased interest in political knowledge.

Most research finds that overall levels of political knowledge held by the public are quite low. Political scientist Philip E. Converse succinctly summed up a vast body of literature when he noted that the variance in political information is very high although the mean level is very low. There is no consensus among scholars as to how much knowledge the public needs to function effectively in a democratic system, but there is agreement that much of the public lacks adequate knowledge to participate effectively. Research has found that there has been relatively little change over time in the overall levels of knowledge held by the public, despite increases in education and changes in the mass media environment that arguably could have raised knowledge levels.

A consistent research finding is that there are dramatic differences in knowledge levels across groups in the population, supporting Converse's observation of variance in levels of political knowledge, with the well-educated more knowledgeable than the less-educated, whites more knowledgeable than racial minorities, men more knowledgeable than women, and middle-aged people more knowledgeable than the young or the very old. These differences apply to most types of political knowledge, but not to all. For example, women are at least as knowledgeable as men, if not more so, on many questions about local government and politics.

Research also confirms that political knowledge and political engagement are highly correlated, with more knowledgeable people far more likely to vote and take part in other political activities. More knowledgeable people also are more likely to vote for candidates who take positions on issues consistent with those of the voter.

Survey researchers have worried that testing respondents' knowledge about politics (or about any other topic) has the potential to harm the rapport between the interviewer and the respondent and even lead respondents to terminate the interview. There is little solid evidence that this is a serious problem, and anecdotal evidence suggests that response rates for surveys containing knowledge questions are not lower than for other surveys.

How best to measure political knowledge has been the subject of considerable debate among survey researchers. Assuming that *political knowledge* is defined as factual information about politics retained in memory, asking factual questions of a random sample of the population of interest would be the most appropriate method. But which questions? How many? What kind? Should respondents guess if they are unsure? Scholars have relied upon the experience of educational testing for guidance on many of these issues, such as the idea that tests of knowledge should include many questions, cover most or all important subdomains of knowledge, and have varying levels of difficulty.

It is generally agreed that a larger number of questions can provide more reliable measurement, whether of knowledge or opinion. But practical considerations in most surveys limit the space that can be devoted to knowledge measures, and thus careful selection of items that discriminate well is essential. Fortunately, considerable research indicates that general political knowledge is not a particularly multidimensional concept, in that different kinds of knowledge—such as knowledge of the institutions and processes of government, the candidates and parties, and the major issues of the day—tend to be highly correlated with one another. This greatly simplifies the survey researcher's task, allowing a relatively smaller number of questions to produce knowledge scales of acceptable reliability as assessed through measures such as Cronbach's alpha.

Scott Keeter

See also Cronbach's Alpha; Election Polls; Knowledge Question; Media Polls; Public Opinion

Further Readings

Converse, P. E. (2000). Assessing the capacity of mass electorates. *Annual Review of Political Science*, *3*, 331–353.

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans know about politics and why it matters.* New Haven, CT: Yale University Press.

Mondak, J. J. (2001). Developing valid knowledge scales. *American Journal of Political Science*, *45*(1), 224–238.

Price, V. (1999). Political information. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of political attitudes* (pp. 591–640). New York: Academic Press.

Zaller, J. R. (1992). *The nature and origins of mass opinion.* New York: Cambridge University Press.

Poll

Poll is a term commonly used to refer to a public opinion survey in which the main purpose is to collect information concerning people's opinions, preferences, perceptions, attitudes, and evaluations of public opinion issues. The performance of public officials, confidence in public and private institutions, enacted policies on topics such as poverty, education, taxes, immigration, public safety, the war on terror, same-sex marriage, abortion, or gun control are only some of the topics that polls typically focus upon. Also, the term *poll* is associated with the measurement of current presidential and congressional vote intention, using either candidates' actual names or generic references such as "Democrat" or "Republican" in order to forecast election results.

A poll is conducted on the basis of sampling principles; that is, a representative group of persons from a specific population is interviewed in order to generalize results to the whole population. A common practice in a poll is to establish an approximate *confidence interval*, also known as *sampling error* or *margin of error*. For instance, a poll may be reported stating that "with 95% confidence the margin of error is plus/minus 3% around the estimated percentage, given a sample size of 1,000 cases." This type of statement means that if the poll were taken repeatedly with different samples of the same size, one might expect that 95% of the time (i.e., 19 times out of 20) the poll estimate would be within the confidence interval.

A poll may be conducted using self-administered or interviewer-administered methods or a combination of both, depending on the survey aims as well as budget constraints. Telephone and face-to-face modes are the predominant interviewer-administered methods used for public opinion polling. Mailing, computer- and Web-based methods are self-administered modes that are also used. Such interviewing modes are used either because of financial reasons, time constraints, or at times when public issues are highly controversial or socially (un)desirable. By removing the interviewer's presence, a pollster may elicit more truthful answers; however, having no live interviewers may considerably reduce the willingness of a respondent to participate.

Question wording plays a critical role in a poll; variations in the question stem may achieve different results. For instance, a question asking, *Would you vote or not for making marijuana legally available for doctors to prescribe in order to reduce pain and suffering?* may yield very different results than a question worded, *Do you support legalizing marijuana for medical use?* Question order is also an important feature in opinion polling, because respondents tend to answer polling questions within the context in which they are asked; hence, the content of preceding questions may affect subsequent answers.

Mostly, polling questionnaires have closed-ended questions with either ordered response scales or categorical response options. Sometimes polls also include a few open-ended questions, which usually are follow-up questions from previous closed-ended questions. Thus, respondents are provided statements or alternatives among which they can choose, as simple as a "Yes/No" format and as extensive as a numeric 11-point scale ranging from 0 to 10 and anchored with "Extremely Dissatisfied" and "Extremely Satisfied." Also, differences in the way response options are presented may yield different patterns of response even when the question wording is kept constant. For instance, providing explicit response option categories such as "Don't Know," "No Opinion," or "Neutral" may change the overall distribution of the other response options.

The order of response options also matters in opinion polling, because in some interviewing modes, respondents tend to select the response option listed either in the first or last position—known as a *primacy effect* and *recency effect*. Furthermore, other elements such as field work timing, weighting procedures, and nonresponse may lead to simultaneous publication of apparently contradictory poll results. Overall, polls are useful for gathering public opinion information, but careful attention must be given to sampling and nonsampling errors when data are reported and analyzed.

René Bautista

See also Closed-Ended Question; Confidence Interval; Election Polls; Margin of Error; Media Polls; Nonsampling Error; Open-Ended Question; Primacy Effect; Question Order Effects; Recency Effect; Response Order Effects; Sampling Error; Social Desirability

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Polling Review Board (PRB)

The Polling Review Board (PRB) is part of the National Council on Public Polls (NCPP), an association comprised of survey organizations from academia, the media, market research, and the news, among other industries. The Polling Review Board supports the NCPP's mission by monitoring the conduct and reporting of polls by member and nonmember organizations. Established in 1969, the NCPP has undertaken to set professional standards on polling and to educate the media, politicians, and the general public on polling, reporting poll results, and the interpretation of reported polls. The NCPP publishes principles of survey disclosures and maintains a Speaker's Bureau and an interactive Web site.

The Polling Review Board publishes reports, available on the NCPP's Web site, that distinguish different types of polls and clarify issues of interest to those in and outside of the polling organizations that make up the NCPP's membership. The paragraph that follows details the NCPP's Principles of Disclosure, the guide used by the NCPP and the Polling Review Board to ensure that consumers of surveys are able to understand reported survey results.

> The NCPP's Principles of Disclosure are aimed at ensuring that publicly available survey results disclose methodological information which will enable consumers of surveys to assess the reliability and

validity of the results reported. The Principles of Disclosure are organized according to three levels of reporting. Level One stipulates the information that all reports of survey findings from member organizations must contain if issued for public release. Level Two disclosure includes information which is provided by a member organization in response to a written request for additional information pertaining to reported survey results publicly released by a member organization. Member organizations are encouraged to provide Level Three disclosures, which include the release of raw datasets, the posting of complete survey wordings, question ordering and percentage results for publicly released survey questions, and a note regarding the survey organization's compliance with NCPP Principles of Disclosure.

Beyond these disclosure levels, the NCPP also details in its Principles of Disclosure a review procedure employed by the NCPP officers and the Committee on Disclosure in which a question is raised regarding member compliance with NCPP Principles of Disclosure and/or where an individual or organization questions the survey methods employed by a member organization's publicly available survey. The Principles of Disclosure were last revised in 2006.

Traci Lynne Nelson

See also Call-In Polls; Deliberative Poll; Election Polls; Exit Polls; Gallup Poll; Internet Pop-Up Polls; Log-In Polls; Media Polls; National Council on Public Polls (NCPP); Poll; Push Polls; Straw Polls; Tracking Polls

Further Readings

Polling Review Board. (2003, March). *Reporting international surveys*. National Council on Public Polls. Retrieved August 22, 2007, from http://www.ncpp.org/?q=node/30

Polling Review Board. (2004, September). *Finding the elusive voter*. National Council on Public Polls. Retrieved August 22, 2007, from http://www.ncpp.org/?q=node/28

Pollster

A pollster is a person who measures public attitudes by conducting opinion polls. Pollsters design, conduct, and analyze surveys to ascertain public views on various subjects. Pollsters typically conduct this work on behalf of clients, including corporations, news organizations, and candidates for public office.

Time magazine first used the term in May 1939, referring to "Dr. George Horace Gallup, punditical pollster of public opinion." But the term appeared only rarely in print until after the infamous presidential Election Night ordeal of 1948 sent the polling profession temporarily reeling. The three major pollsters of the day—Gallup, Elmo Roper, and Alfred M. Crossley—each forecast an Election Day sweep for New York Gov. Thomas E. Dewey, so when President Harry S. Truman won, the *Detroit Free Press* ran the box-score "Truman 304, Pollsters 0."

In the year before the 1948 debacle, there were 15 references to pollsters in *The Washington Post* and *The New York Times*. The Dewey mishap helped put the term into more popular usage; those papers used the term 139 times during the next year. (There were nearly 500 such mentions in 2006.)

After 1948, Gallup and others updated their methodologies, employing probability samples in their estimates. But many had already begun to use the word *pollsters* pejoratively. Lindsay Rogers, a professor of public law at Columbia University, wrote *The Pollsters*, a scathing book, hoping to equate the new profession with a bunch of "hucksters."

For some, the negative label stuck. In 1976, Gallup called the term "sort of denigrating." Roper never used the title, also insisting that his work be called "surveys," not polls. Richard Wirthlin, a Republican consultant, once said the term made him "shudder."

However, *pollster* regained its more neutral meaning on the heels of pollsters' successful implementation of new opinion-gathering techniques following the famous early failure. "Pollsters Bask in the Limelight," the AP wire rang in 1960. "The Pollsters Come Through," read a *New York Times* editorial in 1968.

By then, pollsters had become integral to political campaigns. President John F. Kennedy was the first to hire a pollster when he commissioned Louis Harris to conduct private campaign polls. (Columnist Stewart Alsop called Harris a "pulse-feeler.") President Bill Clinton titled his advisor Stanley Greenberg "Pollster to the President," and a popular movie during the Clinton era, *The American President*, included prominent roles for the actress and actors playing the White House's public opinion advisors and pollster. And in 2004, the presidential candidates spent about $5.5 million on polling, and those running for their party's nomination in 2008 were well on pace to shatter that amount before the first primary votes had been cast.

Most political pollsters also gather opinion data for commercial clients and/or interest groups. Many major media companies and some think tanks also employ pollsters. According to the Bureau of Labor Statistics, which also eschews the term *pollster*, there were approximately 24,000 people employed as "survey researchers" in 2006.

Despite their prominence in politics and market research, pollsters have been little noted in popular culture. Only one major motion picture features a "poll-taker" as a protagonist. In *Magic Town* (1947), Jimmy Stewart turns around a failing polling operation by finding a town whose demographics perfectly mirror that of the nation's.

And the term still confuses many. Democratic pollster Peter Hart recounts a relative-to-be's glee at hearing there was to be an "upholsterer" in the family. One *Washington Post* pollster was complimented on his "Pulitzer" at a gala.

Jon Cohen

See also Crossley, Archibald; Election Polls; Gallup, George; Media Poll; Poll; Probability Sample; Public Opinion; Public Opinion Research; Roper, Elmo

Further Readings

Gallup, G. H. (1940). *The pulse of democracy: The public opinion poll and how it works.* New York: Simon & Schuster.

Moore, D. (1995). *The superpollsters: How they measure and manipulate public opinion in America.* New York: Four Walls Eight Windows.

Rogers, L. (1949). *The pollsters: Public opinion.* New York: Knopf.

POPULATION

All definitions of the population start with a definition of an individual, the elementary units about which inferences are to be drawn. The population is then the collection or aggregation of the individuals or other elements about which inferences are to be made.

In statistical usage, a *population* is any finite or infinite collection of individual elements. That inferences are to be made to the collection of individuals is implied by this definition. The term *universe* is also used in statistics for infinite or hypothetically infinite set of elements.

In survey usage, a *population* is strictly a finite collection of the units from which information is sought in the survey, with additional specification. The term *universe* is avoided, because it implies the infinite and hypothetical; survey research is materially oriented.

There are dimensions of the definition required for implementation purposes. Survey research also often defines populations that are mixtures of units, often where the different types of units are hierarchically ordered. For example, a survey could be designed to collect data from schoolchildren, teachers, and schools in the same survey. All three types of units are part of the population, an aggregate of different types of units.

A population description in survey usage includes the content, the units, the extent, and a temporal dimension. For example, a survey of student performance in a national measurement of educational attainment could specify the population as all students (content), grouped in classrooms (units), in schools in the United States and its territories (extent), in 2008 (temporal dimension). This population is clearly countable and potentially subject to census rather than survey collection.

A population definition often must be modified to meet operational constraints, leading to several types of populations that must be specified. There are different overlapping terms applied to these populations. A *population of inference* may be constrained by operational considerations to a target population. For example, the resources available to the survey may require the elimination of elements that are costly or impractical to reach. A *target population* may consist of all students (content) in public schools (units) in the 50 states, the District of Columbia, and Puerto Rico (extent), enrolled at the time of data collection during the 2007–2008 school year (temporal).

Further restrictions arise as materials are obtained to select the sample. The collection of materials used to select the sample is referred to as a *frame*, and the aggregate of units in the frame is sometimes referred to as the *frame population*. For example, the frame may include only schools known to be in operation at the beginning of the school year; schools opened during that year are excluded.

Some survey researchers define yet another reduction of the population that accounts for nonresponse. For example, not all students may respond to a request to provide information for the survey; students may be absent repeatedly on survey administration days at a school; or parental permission was not obtained for all students selected for study. A *survey population* is sometimes used to describe this latter collection of individuals, implying all elements that would have responded to the survey request rather than only those in the sample that did respond. In principle, the population of inference, the target population, and the frame population can be delineated and counted outside the framework of the survey. The survey population is delineated within the conduct of the survey, less precisely.

The extent of the population will vary within use of data for a particular survey. Surveys are seldom designed to yield information only for the total population. *Subclasses* are subpopulations within the target population for which separate estimates are prepared. A subclass is thus a portion of the sample for which inferences are to be made to the totality of subclass elements in the population. In other words, a subclass is a population with a different extent, or limits. For example, the survey of educational attainment will undoubtedly present findings for male, female, elementary, and other subgroups of students. Each subgroup is itself a population with a different extent than the overall survey population.

The definition of the term *population* is not standardized in the field. Some authors use different terminology to define each of these groups. For example, many textbooks use the term *target population* for *population of inference* used here. *Sampled population* is used to refer to the target population. Careful attention should be given to the definitions used by the authors of survey documents.

James M. Lepkowski

See also Elements; Frame; Population of Inference; Sampling Frame; Target Population; Universe

Further Readings

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Population of Inference

The population of inference refers to the population (or universe) to which the results from a sample survey are meant to generalize. Surveys are used to study characteristics of, and make generalizations about, populations. There are different terms that are used to describe the population, but the most commonly used is the *target population*, which is a finite set of elements to be studied. However, the term *population of inference* (or *inferential population*) is used more often during the conceptualization stage of research studies and surveys.

The target population is specified with the content of the study elements (e.g., general population, students), the units to which the elements belong (e.g., household, school classroom), the geographic boundaries (e.g., country, state), and the time periods (e.g., month, year). In contrast, the population of inference is rather loosely defined in that the specificities of these design elements are not addressed in detail. However, one may argue that the population of inference is more explicitly specified than the *population of interest*.

To illustrate how these terms are related, consider the following example. A research project may proceed as follows: (a) the population of interest is defined as the general population of the United States; (b) the population of inference for the survey is conceptualized to fulfill the research intent by refining the time scope to the year 2008; (c) the target population is operationalized as the noninstitutional civilian persons residing in the conterminous United States between March 1, 2008, and September 30, 2008, who speak English or Spanish. Notice that the definition of the population becomes more crystallized, the breadth of the study population becomes narrower, and the population itself becomes more operationalizable through this process. Of course, these three "populations" may coincide in some cases, for instance, when studying members of certain associations or employees of certain corporations.

Note also that the term *universe* is also often used as a synonym of *population*. The main difference between the universe in statistics and the population considered in surveys is that the former contains a hypothetical infinite set of elements generated by a theoretical statistical model, while the latter is more tangible, as it is a finite set of elements existing in real life. Because surveys are operationalized on the concept of finite populations, the term *universe* is not frequently used in survey research and statistics literature.

Although the conceptualization of each survey is carried out for one specific population of inference and the data collection is conducted on one target population, there may be multiple populations of inference when collected survey data are used. For instance, the

Current Population Survey (CPS) in the United States is conducted to provide official statistics for the employment and economic situation of the U.S. population. However, individual studies using the CPS data make inferences not only about the general population but also its various subgroups, such as female workers, senior citizens ages 65 and older, households living under the U.S. federal poverty threshold, non-Hispanic whites, rural residents, and so on.

Sunghee Lee

See also Current Population Survey (CPS); Finite Population; Frame; Population; Population of Interest; Sampling Frame; Target Population; Universe

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model-assisted sampling*. New York: Springer.

Population of Interest

Most scientific research has some specific groups of interest and attempts to make generalizations about the characteristics of those groups. This is what is termed the *population of interest*. For example, a public health study assesses medical needs among senior citizens; an educational study examines the relationship between high school students' academic performance and their parents' academic attainment; and a marine biology project attempts to investigate the life cycle of humpback whales. The population of interest in the first study is the senior citizens; the second high school students; and the third humpback whales. The same applies to applied social science studies that employ surveys.

While closely related to one another, the population of interest is more loosely defined than the *population of inference* and the *target population*. In fact, the definition of the population of interest is too loose to be directly implemented in survey data collection. In the case of the senior citizen study, who constitutes the senior citizens? Is there any age criterion? What is the reference time period for the survey? The age criterion applied to the scope of senior citizens yields different sets of people depending on the selected ages and time period chosen. Are all U.S. senior citizens eligible for the survey? What about U.S. senior citizens living abroad at the time of the survey? What about those living in nursing homes? These kinds of questions must be answered in order to make the survey operationalizable. The answers to these questions narrow the population of interest to the population of inference and then narrow it further to the target population.

Strictly speaking, survey results cannot be generalized to the population of interest unless it is perfectly aligned with the target population. This, however, is not necessarily practiced in reality. For example, the target population of the General Social Survey (GSS) in the United States is defined as "all non-institutionalized English-speaking persons 18 years of age or older living in the U.S." Most studies based on the GSS data use the results to explain behaviors and attitudes of "American adults," which is different from the target population. This type of misalignment is common, reflecting the gap between the ideal and practical study settings.

Surveys are conducted using fixed resources, so often the budget will not allow for a survey design that reaches every person in the population of interest. It is costly to develop questionnaires in languages other than English and hire and train interviewers for non-English interviews, and it may not be the most effective way to allocate resources given that the proportion of non-English speakers in the general population is not large. Also, even in the unlikely case where there were essentially "unlimited" resources, one may not have access to all of the population elements. For example, it may be impractical, if not impossible, to try to reach prison and jail inmates for general population surveys, or U.S. residents abroad. Errors coming from these factors, however, often do not hamper the generalizability of survey findings when the noncovered portion of the population is relatively small.

Sunghee Lee

See also Finite Population; Frame; Population; Population of Inference; Sampling Frame; Target Population

Population Parameter

Population parameters, also termed *population characteristics*, are numerical expressions summarizing various aspects of the entire population. One common example is the population mean,

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N},$$

where Y_i is some characteristic of interest observed from the element i in the population of size N. Means, medians, proportions, and totals may be classified as descriptive parameters, while there are parameters measuring relationships, such as differences in descriptive parameters, correlation, and regression coefficients.

Although population parameters are sometimes considered unobservable, they are taken to be fixed and potentially measurable quantities using survey statistics. This is because sampling statistics are developed for well-specified finite populations that social science studies attempt to examine and that the population parameters depend on all elements in the population. Before any sort of data collection, population parameters actually are not known. When a census is conducted, all members of the population are observed (in theory), and the "exact value" of the population parameters becomes obtainable. And, by default, the measures taken by a census come to define what is the population "parameter," even if the census is not likely to be exactly accurate. In reality, however, the census is a special case and is not a feasible option for measuring most population parameters.

Instead, social science studies use samples drawn from the population of interest. The population parameters are estimated using estimators. One example is the mean of the sample elements,

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n},$$

where y_j is the same characteristics described previously but measured on the element j in the sample of size n. The sample mean, $\bar{y}$, is used as an estimator of the population parameter, $\bar{Y}$, and a sample mean calculated from a particular sample is an estimate of the population parameters. The key feature of the sample statistics is their representativeness (or unbiasedness) of the population parameters, which is generated mathematically via a probability sample design. Because the sample elements are selected under a random mechanism in probability sampling, $E(\bar{y}) = \bar{Y}$ is ensured, in theory.

It also should be noted that the sample statistics themselves are a random variable with a probability or sampling distribution and are dependent upon the sample design and the realized sample. Unlike population parameters that are constant, estimates from a particular sample may be different from those of another sample drawn from the same population, due to sampling variance and simply because a different set of sample units is selected. This is related to the fact that the unbiasedness of the sampling statistic (e.g., $E(\bar{y}) = \bar{Y}$) is a property of the entire sampling distribution, not of a particular sample. Standard errors associated with the sample estimates measure this sampling variability.

A census provides parameter estimation without sampling errors, but it does not automatically imply that the parameters are measured without error, as the quantities calculated from census data are still subject to nonsampling errors, namely coverage, nonresponse, and/or measurement errors. For example, population values calculated from census data collected using inadequately trained interviewers and plagued by low response rates well may have larger errors than sample estimates from well-designed and administered survey data. In many cases, resources used for a census can be redistributed for a sample survey with better quality control—especially with respect to nonresponse and measurement errors. The sample estimates for population parameters are likely to be more accurate than the population values from a census. Apart from the cost issue, this is another reason why sample surveys are more widely used to study population parameters rather than a census.

Sunghee Lee

See also Finite Population; Population; Population of Inference; Survey; Target Population

Further Readings

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.
Kish, L. (1965). *Survey sampling.* New York: Wiley.
Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model-assisted sampling.* New York: Springer.

Positivity Bias

Positivity bias refers to the phenomena when the public evaluates individuals positively even when they have negative evaluations of the group to which that individual belongs. It is commonly seen within the political science literature that examines positive respondent evaluations of individual political leaders in spite of that respondent's negative views on government in general. This phenomenon has been seen for more than 70 years.

It has been suggested that the public will generally evaluate specific individuals more favorably than impersonal objects or groups. Poll results evaluating political leaders suggest that this positivity bias can be found regardless of the leader's party, ideology, or relative fame. The positivity bias has also been seen in evaluations of individuals in careers as wide ranging as Hollywood actors and actresses to athletes to union leaders.

This phenomenon seems to exist regardless of the amount or quality of information a respondent has for the object of evaluation. Studies indicate that the positivity bias is even seen when the respondent has never heard of the individual he or she is evaluating. In this situation, some suggest that the respondent assumes a favorable evaluation because he or she presumes that the person must be famous in order to be included in the survey. Research has also revealed that a positivity bias can be found in any subpopulation categorized by demographics like age, gender, race, and religion.

Examples

Within public opinion research, there are many examples of positivity bias to draw from. For instance, during the Watts riots in Los Angeles in August 1965, a survey of African Americans revealed that only 2 of the 13 political leaders assessed received negative evaluations. This was unexpected since the riots were in response to racially motivated brutality by the police department. Although evaluations of the police department and city government were low, the political leaders were generally held in relatively high regard.

Another example comes from surveys conducted in the middle of the Watergate scandal in the 1970s. Even in the midst of a scandal, only 4 of the 18 leaders assessed received negative evaluations. Once again, the public generally held the government in relatively low regard, yet 77% of the leaders evaluated received positive evaluations.

Positivity bias has a long history of scholarly research in the social sciences. It is still common to see published research examining these phenomena, but it is rare to see it explicitly referred to as *positivity bias*. This phenomenon has drawn the interest of a host of scholars seeking to understand why the public loves their congressional representatives, yet hates Congress.

Source

The source of the positivity bias is debatable. Within psychology, some believe that these biases are a part of the consistency paradigm, while others suggest it is more closely associated with the person perception literature. Political scientists tend to think of it in terms of perceived similarity, such that respondents believe the individuals they are evaluating are more similar at the individual level than the group to which the individual belongs.

Research has also attempted to determine whether the positivity bias is an artifact of the survey instrument. In a battery of experimental tests, research has shown that this bias is not associated with a respondent's desire to please the interviewer regardless of the instrument used. Another possible explanation is the way these evaluations are measured. These types of questions typically use a Likert-based, bipolar scale. Research has been conducted to determine if measurement options explain the positivity bias. Studies also sought to determine if the use of the subject's official title influenced a respondent's evaluation. In the end, results consistently suggest that the survey instrument is not responsible.

James W. Stoutenborough

See also Approval Rating; Bipolar Scale; Demographic Measure; Likert Scale; Respondent; Social Desirability

Further Readings

Lau, R. R., Sears, D. O., & Centers, R. (1979). The "positivity bias" in evaluations of public figures: Evidence against instrument artifacts. *Public Opinion Quarterly*, *43*, 347–358.

Sniderman, P. M., Brody, R. A., Siegel, J. W., & Tannenbaum, P. H. (1982). Evaluative bias and issue proximity. *Political Behavior*, *4*, 115–131.

Post-Stratification

Stratification is a well-known sampling tool built on the premise that like units in a population should be treated similarly. It is a statistical fact that grouping similar units, when sampling, can generally reduce the variance of the survey estimates obtained. Stratification can be done when selecting units for study, or it can be carried out afterward. The latter application is usually termed *post-stratification*.

Illustration

To illustrate the differences between stratification and post-stratification, assume a researcher is interested in the total poundage of a population that consisted of 10,000 baby mice and one adult elephant. Suppose, further, that the average baby mouse weighted 0.2 pounds but the elephant weighted three and one-half tons or 7,000 pounds. This would mean, if the whole population were to be enumerated, that the researcher would obtain a total of

$$(10{,}000) \times (0.2) + (1) \times (1) \times (7000) \\ = 2000 + 7000 = 9{,}000 \text{ pounds.}$$

Now, if the researcher drew a sample of size $n = 2$ from this population, she or he would not get a very good estimate of the poundage, unless she or he took the elephant as one of the selections. So, naturally the researcher would stratify, taking the one elephant plus one of the mice at random.

If the researcher took account of population sizes, then she or he would multiply the poundage of the mouse selected by $N_1 = 10{,}000$ and add it to the poundage of the $N_2 = 1$ elephant, and the estimated total poundage over repeated samples would be 9,000 pounds (as shown in the preceding formula). Of course the individual mice vary in size (with a standard error of 0.005 pounds, say), so the researcher would not expect to hit the total "dead on" each time, but she or he might come very close, even with this small a sample (i.e., 9,000 pounds on the average would be the estimate with a standard error of 50 pounds).

How does this example change if the researcher post-stratified? Put another way, what if the researcher decided to stratify *after*, *not before*, she or he had selected the sample of $n = 2$? Suppose, to be specific, that the researcher had taken a simple random sample without separate strata for elephants and mice?

Well, first of all, of the $(10{,}001) \times (10{,}000)/2$ or approximately 50 million samples of two elements, only 10,000 will have exactly one elephant and one mouse. All the other samples will have two mice, and there would be no way to get a good estimate from these samples no matter what the researcher did—a big price to pay for not stratifying before selection. To be specific, if two mice are selected, the expected estimate is

$$(10{,}001/2) \times (0.2) + (10{,}001/2) \times (0.2) = 2{,}000.2.$$

The remaining 10,000 samples, with one elephant and one mouse, do even more poorly unaided. For them the researcher will have an unadjusted expected estimate of

$$(10{,}001/2) \times (0.2) + (10{,}001/2) \\ \times (7{,}000) = 35{,}004{,}500.1.$$

This second result, however, can be "saved" by post-stratification in the same way as was previously illustrated for stratification, since each sample has one mouse and one elephant. The calculations here are

$$[10{,}000/(10{,}001/2)]/(10{,}001/2) \times (7{,}000) \\ + [1/(10{,}001/2)]/(10{,}001/2) \times (.2) = 9{,}000.$$

In this case, the researcher gets back what the stratified estimator provided, but there is clearly a big risk that she or he might get a sample that would not be usable (i.e., that post-stratification cannot save).

Implications of the Illustration

Sometimes even a simple illustration like this makes clear some of the general issues. Three will be mentioned:

1. Stratification and post-stratification can often give the same results.
2. If a researcher can stratify beforehand, usually it is wise to do so. That way she or he can better control the sample sizes in each stratum and guarantee any efficiency gains that the stratification may achieve.
3. Stratification can be used to oversample subsets of a population. If post-stratification is used, the sample sizes cannot be enlarged over what they would have been in its absence.

Generalizations

Both stratification and post-stratification can be used with designs that run the gamut from a simple random sample to a complex multi-stage one. All that is needed is to have a usable population total Y that can be estimated from the sample, say, by y. For each such pair

(Y, y) to post-stratify, a researcher simply adjusts or post-stratifies the original sample estimates by factors of the form $w = Y/y$. In the previous illustration they were

$$Y_1/y_1 = [10{,}000/(10{,}001/2)] \text{ and}$$
$$Y_2/y_2 = [1/(10{,}001/2)].$$

When the application of stratification and post-stratification were first introduced in a data-poor and information-poor world, there were very few known or assumed-to-be-known values Y. But, as early as the 1940s, the problem of trying to handle multiple (Y, y) pairs arose. For example, suppose a researcher wanted to post-stratify an area probability household sample by age, gender, race, and state. The Current Population Survey (CPS) comes to mind. Suppose further that the researcher only had, as is typical, usable marginal totals for some of these values, not a complete cross-classification of all of them. Raking, or raking ratio estimation, was invented for situations like this.

Raking is an extension of standard post-stratification. It works as follows. Under regularity constraints, raking operates in the same way as post-stratification, but iteratively. First, one post-stratifies on one set of marginals (say, age); then, using the newly adjusted data, one post-stratifies again, but this time on a second set of marginals (say, gender); and so forth. Methods differ here, but most practitioners of raking cycle back to the beginning and repeat the process until they have achieved the desired degree of closeness. When W. Edwards Deming and F. F. Stephan first invented the method, "computers" were still human beings, so only simple problems could be handled. Today, there have been examples of raking in which up to 60 or more marginals have been employed.

Issues with raking arise when trying to employ many marginals. Most important among these is that the process may not converge, or, if it does, there can be an attenuation of the post-stratified weights. One standard diagnostic employed in raking or even simpler forms of post-stratification is to calculate the relative variance of the weights themselves, before and after adjustment. As the relative weight variance grows, the possibility may exist that the sample is not very representative of the population or that it is simply too small to simultaneously fit all the population marginals required of it. Standard advice in post-stratification would be to collapse marginals with effective sample sizes of less than 20 or 30.

Post-stratification sometimes can only be done after the data are collected—for example, as in the CPS example, stratifying by gender is best done after selection. The use of post-stratification to better align a sample with known outside information can have three distinct goals:

1. *Reducing sampling error*. The implication in some sampling texts is that variance reduction is the key reason post-stratification is used. In most cases although this is valuable, it is the second and third goals that predominate.
2. *Handling differential nonresponse*. While variance reduction can arise here too, usually it is bias reduction that is the goal and sometimes comes at the expense of variances.
3. *Adjusting for undercoverage*. This type of adjustment might be done to bring surveys from incomplete frames up to some common level. One example might be post-stratifying interviews obtained by using random-digit dialing selections from a frame of landline telephone numbers that was brought up to age, gender, and race totals from the CPS (which uses a frame that includes households without landlines).

Practitioners often use post-stratification to achieve several of these objectives at the same time. The variety of approaches used goes well beyond what is discussed here. Good case studies are found in the survey metadata systems of major government programs sponsored by national statistical offices. Applications by the major polling organizations also should be sought out, as these are usually good examples.

Fritz Scheuren

See also Current Population Survey (CPS); Post-Survey Adjustments; Raking; Strata; Stratified Sampling

Further Readings

Cochran, W. (1977). *Sampling techniques*. New York: Wiley.

Deming, W. E. (1942). *The statistical adjustment of data*. New York: Dover.

Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample survey methods and theory* (Reprinted ed., 2 vols.). New York: Wiley.

Lohr, S. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury.

POST-SURVEY ADJUSTMENTS

Post-survey adjustments refer to a series of statistical adjustments applied to survey data prior to data analysis and dissemination. Although no universal definition exits, post-survey adjustments typically include data editing, missing data imputation, weighting adjustments, and disclosure limitation procedures.

Data Editing

Data editing may be defined as procedures for detecting and correcting errors in so-called raw survey data. Data editing may occur during data collection if the interviewer identifies obvious errors in the survey responses. As a component of post-survey adjustments, data editing involves more elaborate, systematic, and automated statistical checks performed by computers. Survey organizations and government statistical agencies may maintain special statistical software to implement data editing procedures.

Data editing begins by specifying a set of *editing rules* for a given editing task. An editing program is then designed and applied to the survey data to identify and correct various errors. First, missing data and "not applicable" responses are properly coded, based on the structure of the survey instrument. Second, range checks are performed on all relevant variables to verify that no invalid (out-of-range) responses are present. Out-of-range data are subject to further review and possible correction. Third, consistency checks are done to ensure that the responses to two or more data items are not in contradiction. In computerized data collection, such as Computer-Assisted Telephone Interviewing (CATI), Computer-Assisted Personal Interviewing (CAPI), or Web surveys, real-time data editing is typically built into the data collection system so that the validity of the data is evaluated as the data are collected.

Missing Data Imputation and Weighting Adjustments

Imputation and weighting adjustments are standard tools for dealing with missing data in surveys. Missing data result from two types of nonresponse: unit nonresponse and item nonresponse. *Unit nonresponse* occurs when no data are collected from a sampled unit, while *item nonresponse* occurs when no data are obtained for some items from a responding unit. In general, imputation is employed for item nonresponse, while weighting is reserved for unit nonresponse.

Imputation is the substitution of missing data with estimated values. Imputation produces a complete, rectangular data matrix that can support analyses where missing values might otherwise constrain what can be done. The statistical goal of imputation is to reduce the potential bias of survey estimates due to item nonresponse, which can only be achieved to the extent that the missing data mechanism is correctly identified and modeled.

Many imputation techniques are used in practice. *Logical imputation* or *deductive imputation* is used when a missing response can be logically inferred or deduced with certainty from other responses provided by the respondent. Logical imputation is preferred over other imputation methods because of its deterministic nature. Hot-deck imputation fills in missing data using responses from other respondents (donor records) that are considered similar to the respondents' missing data with respect to some characteristics. In cold-deck imputation, the donor may be from a data source external to the current survey. In regression imputation, a regression model is fitted for the variable with missing data and then used to predict the missing responses.

Both deck imputation and regression imputation could lead to underestimation of the variance, because the imputed values are either restricted to those that have been observed or they tend to concentrate at the center of the distribution. The multiple imputation method provides a valuable alternative imputation strategy because it supports statistical inferences that reflect the uncertainty due to imputation. Instead of filling in a single value for each missing response, the multiple imputation method replaces each missing value with a set of plausible values. The multiply imputed data sets are then analyzed together to account for the additional variance introduced by imputation.

Weighting adjustments increase the weight of the respondents to compensate for the nonrespondents. Weighting adjustments typically begin with the calculation of the *base weight* to account for the sample design. The base weight, defined as the reciprocal of the probability of selection, accounts for the unsampled cases in the sampling frame. Various adjustments may be made to the base weight to derive the final analysis weight for the respondents. In random-digit dial (RDD) surveys, for example, the base weight may be adjusted for multiple telephone lines, nonresolution of released telephone numbers, screener nonresponse, eligibility rate, within-household selection probability, nonresponse, and so on.

Many sample surveys target a specific subpopulation. If the survey fails to determine the eligibility status for a proportion of the sample, an eligibility adjustment may be made to the base weight by distributing the base weight associated with cases of unknown eligibility status to cases for which the eligibility status is known.

Nonresponse adjustments are typically done in most surveys to compensate for eligible nonrespondents. In weighting class nonresponse adjustments, respondents and nonrespondents are classified into weighting classes based on pertinent information available for both groups. Within each weighting class, the weight of each respondent is increased by a factor equal to the inverse of the response rate in the weighting class. The weighting classes are typically defined by variables that are correlated to response propensity as well as the variables of analytical interest. Response propensity scores are sometimes used to define weighting classes or to derive the adjustment factor directly.

Weighting adjustments often involve additional steps to adjust for discrepancies between the sample and population distributions. Raking ratio adjustments make the marginal distributions of the weighted sample conform to those of the population. When the joint distribution of some population characteristics is known, post-stratification adjustments may be carried out to ensure that the joint distribution of the weighted sample match that of the population.

Disclosure Limitation Procedures

Government agencies and their contractors often are required by law or established policies to protect the confidentiality of the respondents if the survey data are to be released to the public in the form of microdata files or statistical tables. The required protection is achieved by the application of statistical disclosure limitation procedures to the survey data before public release. Specific disclosure limitation procedures depend on the nature of the data to be released. Common procedures include suppression, removing identifiers, sampling, swapping or switching, rounding, adding random noise, top or bottom coding, generation of simulated or synthetic data, and so on.

Y. Michael Yang

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Disclosure Limitation; Hot-Deck Imputation; Imputation; Item Nonresponse; Multiple Imputation; Post-Stratification; Raking; Raw Data; Response Propensity; Suppression; Unit Nonresponse; Weighting

Further Readings

Federal Committee on Statistical Methodology. (1994). *Report on statistical disclosure limitation methodology*. Statistical Working Paper 22. Washington DC: Office of Management and Budget.

Kalton, G. (1983). *Compensating for missing survey data.* Ann Arbor: University of Michigan, Institute for Social Research, Survey Research Center.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Precision

Precision in statistical surveys relates to the variation of a survey estimator for a population parameter that is attributable to having sampled a portion of the full population of interest using a specific probability-based sampling design. It refers to the size of deviations from a survey estimate (i.e., a survey statistic, such as a mean or percentage) that occurs over repeated application of the same sampling procedures using the same sampling frame and sample size. While *precision* is a measure of the variation among survey estimates, over repeated application of the same sampling procedures, *accuracy* is a measure of the difference between the survey estimate and the true value of a population parameter. Precision and accuracy measure two dimensions for a survey estimate. A sampling design can result in survey estimates with a high level of precision and accuracy (the ideal). In general, sampling designs are developed to achieve an acceptable balance of accuracy and precision.

The sampling variance is a commonly used measure of precision. Precision can also be represented by the standard error (the square root of the sampling variance for a survey estimate), the relative standard error (the standard error scaled by the survey estimate), the confidence interval width for a survey estimate (using the standard error and the values from an assumed probability density function for the survey estimate), and difference between estimated percentiles for a population parameter (for example, the intraquartile range for the survey estimate).

The sampling variance is conceptually the squared difference between the estimate for a specific sample and the expected value of the estimate summed over all possible samples selected in the same fashion with the same sample size using the same sampling scheme. The sampling variance is different from the classical "population" variance (a measure of the variation among the observations in the population) in the sense that the population variance is a constant, independent of any sampling issues, while the sampling variance becomes smaller as the sample size increases. The sampling variance is zero when the full population is observed, as in a census.

Because only a single sample is selected and the expected value of an estimate is unknown, sampling theory has provided formulae to estimate the sampling variance from a single sample. To allow for the computation of a sampling variance, the sample scheme must be reproducible, that is, it can be completely replicated. Moreover, to compute an unbiased estimate of the sampling variance, (a) every unit in a sampling frame needs to have a positive chance of being selected (the unit selection probability is greater than zero), and (b) every pair of units must have a positive chance of being in a sample (the joint selection probability for any pair of units is greater than zero).

The sampling variance is a function of the form of the statistic, the underlying population, and the nature of the sampling design, and it is called the "design-based sampling variance." The design-based variance generally assumes the use of a sampling weight, which is computed from the inverse of the probability of selection of the sample.

A statistical sample survey has two general forms of statistics: (a) linear combinations of the survey data (for example, a total) and (b) nonlinear combinations of the survey data. Nonlinear combinations include the ratio of two estimates (for example, a mean or a proportion in which both the numerator and the denominator is estimated) and more complex combinations such as regression coefficients. For linear estimates with simple sample designs (such as a stratified or unstratified simple random sample) or complex designs (such as stratified multi-stage designs), explicit equations are available to compute the sampling variance. For the more common nonlinear estimates with simple or complex sample designs, explicit equations are not generally available, and various approximations or computational algorithms are used to provide an essentially unbiased estimate of the sampling variance.

Two primary forms of sampling variance estimators are available for complex sample designs: (a) the procedures based on the Taylor series linearization of the nonlinear estimator, using explicit sampling variance equations, and (b) the procedures based on forming pseudo-replications of the sample or pseudo-replicates of the survey estimates. The first method uses the classical Taylor series linearization of a nonlinear estimator to construct an analytic variable representing the first-order linearized form of the survey estimator using weighted totals and the data from individual sample members. This linearized analytic variable is used in the explicit sampling variance equations to estimate the sampling variance. Within the class of pseudo-replication procedures, the balanced repeated replication (BRR) procedure, the jackknife procedure, and the bootstrap procedure are most widely used or discussed. The pseudo-replication procedures compute a number of survey estimates using partial subsamples selected from the single sample. The sampling variance is then estimated as a function of the variance among these pseudo-replicated survey estimates. Both forms of sampling variance estimators generate similar estimates of the sampling variance, and the use of one form may be more desirable for specific survey estimators and situations.

The precision for survey estimates is a function of number components in a survey. In general, sampling designs are developed to achieve a balance of precision relative to the population, the information required, and the available resources for the survey.

Frank Potter

See also Balanced Repeated Replication (BRR); Bootstrapping; Census; Confidence Interval; Jackknife Variance Estimation; Parameter; Population; Probability Sample; Sample Size; Sampling Frame; Sampling Variance; Standard Error; Statistic; Taylor Series Linearization; True Value

PRECISION JOURNALISM

Precision journalism is a term that links the application of social science research methods (including survey research methods) to the practice of gathering information for the news purposes of journalists. Similar to a social scientist, a precision journalist discloses the data collection methodology well enough that another precision journalist or researcher could replicate the research studies and ostensibly would reach the same conclusions.

The term was coined by Everette E. Dennis in 1971 as part of seminar he taught at the University of Oregon. The concept then was explicated by one of his students, Neil Felgenhauer, in a term paper that later became a book chapter. Most of the "new journalism" of the time that inspired Dennis's seminar was the creation of talented writers (e.g., Tom Wolfe) who used fiction techniques to construct powerful narratives about current events. The class discussion compared this semi-fictional approach to journalism with precision journalism techniques. As Dennis and William Rivers noted in a 1974 report, although other journalists are pushing news reporting toward more of an art, precision journalists are pushing it toward more of a science. The origins of precision journalism go back to the first public opinion polls that used systematic sampling methods instead of just gathering readily available person-on-the-street interviews. George Gallup based a newspaper column on his national polls that used rigorous survey methods, which, for example, led to a much more accurate pre-election prediction of the 1936 Landon-Roosevelt election than that of the more well-known and much larger unscientific poll conducted by the *Literary Digest* magazine.

Other early users of precision journalism were the television networks wanting to be first to announce the presidential winner on Election Day, although not all their early usage of these new reporting techniques proved reliable or accurate. For example, in 1960, CBS used a statistical model based on the timing of and results from the pre-election polls they had been conducting. The model captured the candidate standings at given points in time and compared them to the presidential candidate standings at the same points in times in the previous election. The initial findings from the model proved to be incorrect, and the resulting bias produced an incorrect early call for Richard Nixon as the projected winner over John F. Kennedy.

In the 1960s, news coverage of the civil rights and anti-war movements fueled the need for new reporting techniques. Standard journalism traditionally was focused on gathering information from the most visible spokespersons for the respective movements—a top-down approach—and thus tended to place too much importance on what was said by these elite sources, who often had their own personal agendas. In contrast, *Newsweek* magazine commissioned pollster Louis Harris to do special civil rights surveys among black citizenry to uncover a broader and more accurate understanding of the attitudes held by the black community.

In 1967, when there were race riots in Detroit, the Knight Newspapers sent Philip Meyer from their Washington bureau to help *The Detroit Free Press* cover the ongoing story. Meyer stayed in Detroit to conduct a survey of residents in the affected neighborhoods in order to measure the grievances held by blacks and thus the root causes of the riot. (Meyer had learned these research techniques as a Nieman Fellow at Harvard University in the previous academic year.) The news stories that resulted from the surveying were one of several factors that earned the *Free Press* the 1968 Pulitzer Prize for general local reporting. Following this, Meyer was assigned to study Miami to further utilize precision journalism methods to aid news coverage of racial problems there before and after the 1968 assassination of Dr. Martin Luther King, Jr. In 1969, the Russell Sage Foundation sponsored Meyer to take a leave of absence from Knight Newspapers and direct a project to prepare a precision journalism handbook for journalists, which resulted in the publication of the first edition of Meyer's book, *Precision Journalism.*

Sometimes precision journalism is confused with *computer-assisted reporting* (including using the Internet to do information searches), because precision journalism often does involve large-scale data collection and analyses aided by computers. But computers are neither necessary nor sufficient for the conduct of precision journalism. Precision journalism includes myriad social science research methods and is not limited to survey research. Over the years, its practitioners have used content analysis of public records, experimental designs, and other quantitative and qualitative behavioral science methods.

In the mid-1970s, the National Science Foundation helped the acceptance of precision journalism by sponsoring two training programs at Northwestern University for midcareer journalists. Reporters from *The New York Times* and *The Washington Post* participated and followed up with related projects at their home newspapers, which helped introduce precision journalism to many other editors and reporters across the country. By the late 1970s, journalism students at Northwestern's Medill School of Journalism and elsewhere were being taught the methods of precision journalism both at the undergraduate and graduate levels. Subsequently, the National Institute for Computer-Assisted Reporting (NICAR) incorporated precision journalism concepts in its training coursework. By the present decade, the use of precision journalism methods is considered fairly routine for most major and many smaller news organizations.

Paul J. Lavrakas

See also Exit Polls; Gallup, George; Horse Race Journalism; Media Poll

Further Readings

DeFleur, M. H. (1997). *Computer-assisted investigative reporting*. Mahwah, NJ: Lawrence Erlbaum.

Dennis, E. E., & Rivers, W. L. (1974). *Other voices: The new journalism in America.* San Francisco: Canfield Press.

Meyer, P. (2002). *Precision journalism: A reporter's introduction to social science methods* (4th ed.). Lanham, MD: Rowman & Littlefield.

Precoded Question

Precoded questions refer to survey items for which response categories may be identified and defined exhaustively, or very nearly so, prior to data collection activities. Precoding questions involves specifying the coding frame (i.e., the set of possible answers) and associating each response category within the frame with a value label (which is typically, but not necessarily, numeric).

The term *precoded* also refers to a type of question that is asked by interviewers as though it is an open-ended question, but it has precoded responses that interviewers are to use to match (code) respondents' answers rather than copy down the verbatim given by the respondent. This also is referred to as *field coding*.

The use of precoded questions delivers two important benefits to the survey researcher. First, their use minimizes the time to prepare the answers for statistical analysis following the completion of data collection activities. Second, because the data are coded as they are collected, their use is thought to reduce coder variance.

Value Labels

Value labels for precoded questions are typically assigned in a manner that coincides with the measurement level (nominal, ordinal, interval, or ratio) implied by the item in order to aid interpretive analysis of the survey data. For example, value labels assigned to response possibilities that correspond to interval or ratio level measures typically are numerical, with the set of numbered values chosen to reflect the ordered and evenly spaced characteristics assumed by these measurement levels. When ordinal measures are involved, numerals are typically used for the value label codes, and the number values chosen appear in a consecutive sequence that is directionally consistent with the ordinal character of the measure's response categories. (Although the use of alphabetical characters would be equally effective in communicating the directional sequence of the response categories, use of numerals is almost always preferred because manipulation of alphabetical or "string" value labels by statistical analysis software applications typically is far more cumbersome than when numeric value labels are used.) In contrast, *code value labels* for items featuring nominal levels of measurement may be assigned in an arbitrary manner, as they bear no meaning or relationship to the response categories themselves. Therefore, while sequenced numerals or letters may be used for these value labels, these often are assigned in an order corresponding to the sequence in which the response choices are documented in the research instrumentation.

These guidelines are reflected in the self-administered questionnaire examples in Figure 1, shown with the code value labels represented by the small numerals near the check boxes corresponding to the response choices offered.

It is worth observing that, since the value labels for a nominal measure item are arbitrary and meaningless, it often is functional to use alphabetical value labels to aid the analyst's memory in associating the value labels with the code definitions themselves. In the nominal measurement item shown in Figure 1, for example, the letter O (for Own) might have been chosen for the value label instead of 1, R (for Rent) might have been assigned instead of 2, and A (for Another) might have been used instead of 3.

Interval-Level Measurement	**"Not including today, on how many of the past seven days have you read that day's edition of a newspaper?"** 7 ☐ All, Seven Days 6 ☐ Six Days 5 ☐ Five Days 4 ☐ Four Days 3 ☐ Three Days 2 ☐ Two Days 1 ☐ One Day 0 ☐ None, Zero Days
Ordinal-Level Measurement	**Please indicate how strongly you agree or disagree with the following statement: "Looking back on my life, I am fairly well satisfied."** 7 ☐ Completely Agree 6 ☐ Mostly Agree 5 ☐ Slightly Agree 4 ☐ Neither Agree nor Disagree 3 ☐ Slightly Disagree 2 ☐ Mostly Disagree 1 ☐ Completely Disagree
Nominal-Level Measurement	**Do you own your home, rent your home, or do you have some other living arrangement?** 1 ☐ Own home 2 ☐ Rent home 3 ☐ Another living arrangement

Figure 1 Examples of precoded labels for items having alternative levels of measurement

Types

Precoded questions may be either closed-ended (where response categories are specified exhaustively prior to data collection), partly closed-ended (where nearly all response categories can be and are specified prior to data collection, but a response category of "Other (specify):" is included to accommodate unanticipated or unusual responses), or open-ended (where responses are expected to be spontaneously self-generated by the respondent). Still, because of the difficulty of anticipating the responses that respondents will provide to an open-ended question, precoded questions most often are closed-ended or partly closed-ended. Furthermore, researchers often want the richer detail provided by full verbatim responses to an open-ended question rather than losing this detail and having interviewers choose a precoded response.

There are two varieties of open-ended question for which precoding is common, however. The first is the self-coding open-ended question. These are questions in which the form of the respondent's answer is implied and logically restricted by the question itself. *How many books did you read last month?* would be an example of such a question. Logically, a respondent either must provide an answer like "None" (i.e., zero) or some other whole number or must fail to provide an answer by saying "I don't know" or refusing to respond to the item. The whole number answers self-code to that numeric value and the Don't Know and Refused possibilities may be assigned unlikely number values, such as, for example, 88 and 99, respectively.

The second type of precoded open-ended question has response "categories" that are clear within a given culture and thus easily anticipated. A routinely used item asking the respondent to describe his or her formal education, such as *What is the highest grade or level of school you have completed?*, is a good example of this kind of question. In this case, the range of possibilities can easily be anticipated, and the interviewer can be expected to code the respondent's answer reliably as soon as it is given. For example, the following set of response categories and corresponding code value labels might be used by interviewers:

No formal education = 00,

Less than high school graduate = 01,

High school graduate or G.E.D. = 02,

Some college, but no degree = 03,

Two-year college degree (associate's degree) = 04,

Four-year college degree (e.g., B.A. or B.S.) = 05,

Some graduate school = 06,

Master's degree (M.A., M.S., M.B.A., etc.) = 07,

Doctoral degree (M.D., D.O., Ph.D., J.D., etc.) = 08,

Don't know = 88, and

Refused = 99.

Jonathan E. Brill

See also Closed-Ended Question; Coder Variance; Coding; Field Coding; Interval Measure; Level of Measurement; Nominal Measure; Open-Ended Question; Ordinal Measure; Ratio Measure; Verbatim Responses

Predictive Dialing

Predictive dialing is a telephone call placement technology that is used in survey research to improve utilization of interviewer time during computer-assisted telephone interviewing (CATI) surveys. In random-digit dialing (RDD) surveys, typically fewer than 15% of dialings are answered by a human; for the remaining 85%, and when predictive dialing is not used by a telephone research center, interviewers must disposition unanswered numbers, data or fax lines, disconnects, and answering machines. With predictive dialing, the dialer will handle many of these unproductive calls in the "background," passing calls to interviewers only when a call connects with a human being. The resulting increase in the proportion of "talk time" for the interviewer—time spent by the interviewer persuading and/or interviewing respondents—not only provides direct cost savings, with less interviewer time needed in total for the same number of interviews, but may provide an indirect gain because, in theory, interviewers remain more focused and engaged in the core of their work (i.e., speaking with respondents).

Predictive Versus Nonpredictive Dialing

From an operational standpoint, what distinguishes predictive dialing from nonpredictive dialing is the elimination of the 1:1 interviewer-to-call ratio. If one call is placed for one available interviewer, then the interviewer will be idle while the call is placed and the signal returned, which could be for 30 seconds or more. Autodialing (mechanized dialing while still using a 1:1 ratio), along with automatic signal recognition for engaged (busy), fax, disconnected, and unanswered lines, will partially reduce nonproductive time and increase interviewers' talk time. Predictive dialers include these autodialing technologies but also allow for the probability that not all call results will require interviewer intervention, so typically there are a great many more calls being placed than the number of interviewers available to take them, with the dialer handling the calls that do not require interviewer involvement.

Predictive dialing algorithms utilize time-series or queuing methods, and the variables utilized vary by manufacturer: for example, interviewers in the queue, those nearing the end of a call, connect rates, average call length, and so on. Ideally, all predictive dialing systems strive to make the number of connects equal to the number of interviewers available. Variation from this ideal has two possible results: first, there are more connects than interviewers, leading to either *abandoned calls* or "dead air"—and ultimately alienated respondents. Second, there are too few connections, which results in a decrease in a call center's productivity and an increase in its costs. The extent to which the predictive algorithm errs on the side of "wait time" versus increased "abandonment" (i.e., dropping a call that the dialer detects has been answered by a human since no interviewer is available) is determined by setting the *maximum abandonment rate* parameter on the dialer. At its most conservative setting (i.e., "zero abandonment"), the dialing ratio will be 1:1 (in effect, removing the predictive element). Most survey companies operate with the abandonment rate set under 3% (meaning 3 out of every 100 connects will be abandoned). Of note, research has suggested that setting the abandonment rate too high in telephone surveys is counterproductive, as respondents who have had too many prior contact calls abandoned by a predictive dialer may become frustrated, which increases their likelihood to refuse when a "live" interviewer comes onto the line during a subsequent call attempt.

Telemarketing Versus Research Applications

Predictive dialing originated in the telemarketing industry and is still extensively used there; however, most predictive dialers used by survey research companies operate quite differently than their telemarketing counterparts. Specifically, the algorithms used to predict how many telephone numbers to dial do not utilize *interviewer wait time* as a criterion but instead use the quantity of numbers that need to be dialed to yield connects for already waiting interviewers. This difference is significant both in terms of operation and the designed intention to minimize abandoned calls. Obviously, both algorithms result in abandoned calls, but the research approach is much more conservative.

Another significant difference between telemarketing applications and research applications is the "dead air" that respondents often hear after answering their phone to a telemarketing call, which is the delay between the moment a potential respondent answers the telephone and says "Hello" to the time an interviewer is routed the call and can reply to the respondent, which is the signal that a human is actually on the other end of the line. This delay can arise from several sources:

1. *The switching process.* If the dialer sends a case to an interviewer before the dial is made, there is in effect a complete circuit between the interviewer and

the respondent. During any predictive process (even set at zero abandonment), that circuit needs to be established. If an interviewer is already waiting and both hardware and software configured to minimize the delay, the link is established in milliseconds and will be essentially undetectable by the respondent who picks up a handset from the cradle, although it may be just noticeable by the respondent who has a headset already on his or her ear. The more hardware and software intervention there is between the "connect" signal received by the dialer and an actual switch by the dialer of the audio to the interviewer (e.g., if the call is being routed through a separate switch, as may occur with geographically dispersed systems), the longer this connection will take, and the more noticeable the delay will be.

2. *The queuing process.* As an alternative to abandoning a connection, some systems will queue up the additional connections for which an interviewer is not immediately available and then establish the connections with an interviewer as soon as one becomes available. The delay can be short or long, depending on the system; for example, if there are many interviewers and the average connection time is short, then the probability of an interviewer becoming free in the next few seconds will be very high and the average wait quite short. For situations in which longer delays are more common, some systems have provisions for playing a recorded message to the respondent during the wait for an interviewer to become available.

3. *Automatic answering machine detection (AMD).* The proportion of calls to households that are actually picked by a resident as opposed to answered by an answering machine or voicemail system is in the minority and is declining each year, so substantial cost savings can be made if calls connecting to an answering machine or voicemail can be handled by the dialer rather than an interviewer. The problem is that AMD systems rely on the pause after the respondent says "Hello" to distinguish a human from an answering machine (where the greeting is rarely so brief before the pause occurs). The delay required by this (to hear "Hello" and then the telltale pause) is then often compounded by placing the call in a queue to wait for the next available interviewer. In addition to causing delays, AMD has the further disadvantage of being error prone in detection of answering machines, sometimes adding answering machines to the queue waiting for interviewers because the answering machine has been mistaken for a human, and other times, more seriously, mistaking a human for an answering machine and hanging up on the person after keeping her or him waiting several seconds.

There is no doubt that these delays annoy respondents and increase the likelihood of the respondent hanging up before the interviewer can engage him or her. In the telemarketing industry, this is not viewed as such a problem that it outweighs the substantial cost benefits, since the telemarketer's objective is to maximize sales, which is done by maximizing interviewer talk time, and the pool of numbers to churn through is practically endless. In survey research, however, scientific sampling and the need for high response rates means such alienation cannot be risked, and therefore most survey research applications do not use AMD and instead configure their dialers to eliminate as far as possible any remaining dead air, even though doing so reduces the cost savings that could otherwise be delivered by the predictive dialer.

Yet another downside to predictive dialing, unless it is matched with *preview dialing*, is that it makes it impossible for an interviewer to study the call history for a given number or household before she or he engages that household in the next call attempt to try to gain a completed interview. This is especially disadvantaging in refusal conversion attempts, where past history with the household can be very important for the interviewer to prepare her or his best approach to trying to convert a previous refusal. Preview dialing occurs when a case is sent to an interviewer before the dial is made, so the interviewer can review the case history and any pertinent call notes (such as details of any prior refusal) before initiating the dial. Many dialing systems are flexible, allowing modes to be combined in different ways based on previous results with the sample number or household in order to enhance productivity while maintaining excellent survey practice. For example, one could start all numbers off in preview mode and pass on to predictive dialing only those cases that have had no contact at all in for the first n dials (where n is typically in the range 4–8), since only a very small proportion of such cases will ultimately be identified as households. Alternatively, one might start all cases in predictive mode, but as soon as a sample number is identified as a household, all subsequent dials could be assigned to utilize preview mode.

Jenny Kelly and Dale W. Kulp

See also Computer-Assisted Telephone Interviewing (CATI); Do-Not-Call (DNC) Registries; Outbound Calling; Random-Digit Dialing (RDD); Refusal Conversion; Telemarketing

Further Readings

Hansen, S. E. (2008). CATI Sample Management Systems. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 340–358). New York: Wiley.

Kelly, J., Link, M. W., Petty, J., Hobson, K., & Cagney, P. (2008). Establishing a new survey research center. In J. Lepkowski et al. (Eds.), *Advances in telephone survey methodology* (pp. 317–339). New York: Wiley.

Pre-Election Polls

One of the most common and visible applications of survey research is pre-election polling. These polls typically include one or more trial heat questions, which ask respondents how they will vote in the upcoming election, along with other measures of voter knowledge, attitudes, and likely voting behavior. Unlike most polls, the validity of pre-election polls can be assessed by comparing them with actual election outcomes.

Numerous organizations in the United States and around the world conduct pre-election polls, and the number and frequency of such polls has been growing. Many pre-election polls are public, conducted by news organizations, academic institutions, and nonprofit organizations. Many others are privately conducted by partisan party organizations and campaigns to assist in candidates' message development, resource allocation, and overall strategy. The results of most of these private polls are not made public.

Accuracy

Pre-election polls in the United States have a generally good record of accurately forecasting the outcome of elections. According the National Council on Public Polls (NCPP), the average candidate error for national polls in the 2004 presidential election was 0.9 percentage point; in 2000, the average error was 1.1 percentage points. Polls in state-level races were also very accurate. The average candidate error in 2004 across 198 state polls reviewed by the NCPP was 1.7 percentage points.

But inaccurate polls attract a great deal of attention and tend to be remembered for a long time. Modern polling's most spectacular failure occurred in 1948, when public polls incorrectly forecast the defeat of President Harry S Truman by Thomas Dewey. The problems that led to the 1948 polling debacle were soon addressed, and it is highly unlikely that an error of that magnitude could occur again.

History

Election polling has a long history that precedes the development of modern probability sampling. Straw polls of the 19th century were a popular means by which public opinion in elections was gauged, and an early 20th-century magazine, the *Literary Digest*, conducted a very large and widely followed straw poll in several presidential elections through 1936. The *Literary Digest* poll's spectacular failure in the presidential election of 1936, in which it incorrectly forecast the defeat of President Franklin Roosevelt by the Republican candidate, Alf Landon, by a large margin, discredited straw polls. Despite having more than 2,000,000 respondents, the self-selected nature of the mail survey respondents and the fact that the sampling frame was biased toward more affluent voters led to a gross overrepresentation of Republican voters in the *Literary Digest* sample. A Gallup poll based on a sample that more closely conformed to the principles of probability sampling was quite accurate in the 1936 election and helped to affirm the legitimacy of modern polling methods.

Purposes

Conducting pre-election polls entails many decisions, each of which can affect the accuracy of the poll: (a) the timing of the poll, (b) the sampling method, (c) the determination of likely voters, (d) the trial heat question employed, and (e) the choice of other measures to be included in the poll. Pre-election polls are conducted throughout the election cycle but are considered appropriate for forecasting the election outcome only if taken very close to Election Day. But pre-election polls are conducted for many purposes other than forecasting. Some are conducted at the beginning of a campaign to provide information about the public interest in the election, awareness of the potential candidates, voter views on issues including the importance of different issues, and the likely receptivity of the voters to different messages. Polls conducted during the campaign can provide all of this information as well as reactions of the public to the candidates and events in the campaign. Polls in the latter stages of the campaign are used by the candidates and parties in making decisions about how to allocate campaign resources geographically or in targeting different demographic groups.

Methods for Conducting Pre-Election Polls

Most U.S. pre-election polls are conducted by telephone, though a growing number are conducted via Internet. Telephone samples are typically either random-digit dial (RDD) or from lists of registered voters (so-called registration-based sampling [RBS]). Each method has its advantages and disadvantages. RDD theoretically reaches all potential voters who have a landline telephone but also reaches many people who are not citizens, are not registered, or who are uninterested in politics and very unlikely to vote. The exclusion of voters who do not have landlines but may have cell phones is an increasing limitation of most RDD surveys. Most polling organizations do not routinely include cell phone samples as a part of their RDD telephone surveys as of 2008. Cell-only adults constituted 20% of the population at the end of 2007 and also tend to be younger than other adults and more likely to be Hispanic, characteristics associated with lower levels of voter turnout. But the number of cell-only adults is expected to continue to grow, which may affect the accuracy of RDD samples used for making pre-election forecasts.

RBS samples eliminate many of these problems because they include only those individuals who are actually registered and may include useful information, such as past voting history, that can assist pollsters in estimating the likelihood that an individual will vote in a given election. The quality of RBS samples varies from state to state, however, and telephone numbers are often outdated or not included for many records. As a result, polls based on RBS samples can sometimes be biased by overrepresenting the kind of individuals who are willing to provide telephone numbers or who remain living at a single residence for a long period of time.

One special type of pre-election poll is the *tracking poll*. Tracking polls attempt to document changes in the support of candidates by interviewing a sample of voters every day over a few weeks or more. The typical design for a tracking poll interviews a mix of respondents from a new sample called for the first time on a given day and respondents from a sample first called one or two days before. Responses from interviews over a 2- to 4-day period are weighted together, averaged, and reported. While not as rigorous in design as the typical pre-election poll, tracking polls provide a view of changes in the campaign that may be very useful to journalists and campaign professionals.

Challenges

Perhaps the most difficult challenge for pre-election polls is determining which respondents will actually vote. Average turnout of the voting age population in U.S. presidential elections during the past several elections has been a little more than 50%. But in most pre-election polls, significantly higher percentages of respondents say they intend to vote. This difference reflects both biases in the pool of survey respondents and the fact that voting is a socially desirable behavior that is apt to be overreported.

The political views of likely voters and nonvoters often diverge, and thus it is critical for polls to distinguish those who are likely to vote from those who will not. Considerable research has focused on locating the questions that best discriminate likely voters from others. As a result, most organizations use a combination of questions that include measures of voter registration, voting intention (sometimes on a scale from 0 to 10, or a verbal scale), past voting activity, interest in the campaign, and knowledge of how to vote.

The process of identifying likely voters is relatively straightforward. Typically, a scale combining the turnout measures is created and respondents are ranked on the scale. Separately, an estimate of the likely turnout percentage is made for the election, based on an assessment of voter interest in the current election and a comparison with past elections with similar characteristics. That estimated turnout percentage is used to determine how many of the respondents on the survey's turnout scale will be classified as likely voters. For example, if the predicted turnout percentage is 55%, respondents in the top 55% of survey respondents in the distribution of the turnout scale will be classified as likely voters. For elections with lower turnout, the standard will be more stringent.

Pre-election polls can include a very wide range of content, including interest in the election, knowledge and awareness of the candidates, opinions about the candidates on various dimensions, campaign activities engaged in by the respondent (including donations, volunteer work, efforts to persuade others), importance of various issues, opinions on the issues, and of course, likely vote choice.

Pre-election polls can be valuable whenever they are conducted during the campaign cycle. But only those polls conducted close to the election can have a reasonable assurance of accuracy in predicting a winner of a given race. Many voters pay little attention to a campaign until very close to Election Day, and thus polling several weeks or even several days prior to the election can be inaccurate. One of the main reasons for polling's notorious failure in the 1948 presidential election between Harry S Truman and Thomas Dewey was the fact that most organizations stopped polling days or weeks before the election. Many observers believe that Truman's aggressive efforts at the end of the campaign helped him convert enough voters to eke out a victory, and that this shift in sentiment was not picked up by the polls. Similarly, the failure in the public polls to predict Ronald Reagan's substantial 1980 presidential victory was due at least in part to their stopping data collection on the Sunday before the election, when in fact many voters did not make the final decision whether they would vote for Carter or for Reagan until the day of the election. As such, most polling organizations that offer a final poll forecasting the outcome of the election gather their data through the weekend of the election, and sometimes even through the day before the election. In 1996, the accuracy of presidential polls was correlated with how close to the election they polled.

Scott Keeter

See also Election Polls; National Council on Public Polls (NCPP); Gallup, George; Likely Voter; Overreporting; Random-Digit Dialing (RDD); Registration-Based Sampling (RBS); Sampling Frame; Social Desirability; Straw Polls; Tracking Poll; Trial Heat Question

Further Readings

Crespi, I. (1988). *Pre-election polls: Sources of accuracy and error.* New York: Russell Sage Foundation.

Herbst, S. (1993). *Numbered voices: How public opinion has shaped American politics.* Chicago: University of Chicago Press.

Lavrakas, P. J., & Traugott, M. T. (Eds.). (2000). *Election polls, the news media and democracy*. New York: CQ Press.

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Prefix

Understanding how telephone numbers are assigned is very important when you are designing or implementing a telephone sample. Telephone numbers in the United States, Canada, and the Caribbean consist of 10 digits divided into three components. The first three digits are the area code or Numbering Plan Area (NPA). Area codes usually cover large geographic areas. The next three digits represent smaller geographic areas within an area code and are referred to as the *prefix* (NXX). The term *prefix* is also used to refer to the first three digits of the seven-digit local numbers. When the NXX is combined with the final four digits of a phone number, the result is a unique seven-digit local number that is associated with a unique end user within an NXX. In most areas of the United States, local seven-digit dialing of landlines is the norm. Within these areas, the area code component of a 10-digit number need not be dialed, and calls are switched using only the prefix. Although area codes are unique, prefixes are unique only within an area code. Thus, in making a long-distance telephone call or a call to a different area code, the area code is required in order to define a unique prefix. Not all prefixes are available for assignment of local numbers. Some are reserved for other uses, such as directory assistance (555) and special services such as 911 and 411.

Although the terms *prefix, wire center, rate center, central office*, and *exchange* are frequently used interchangeably, there are distinctions. The wire center, rate center, or central office is the building containing the telephone equipment and switches where individual telephone lines are connected and through which calls to and from individual local numbers are routed. The term *exchange* usually refers to the geographic area served by a particular rate center or wire center. Thus an exchange can be serviced by multiple prefixes (NXX codes). In some areas of the United States, these prefixes might belong to different area codes.

Each prefix contains 10,000 possible local numbers (0000–9999). Within prefixes, local telephone companies assign four-digit local telephone numbers. During the past 20 years the demand for telephone numbers has dramatically increased, driven by technological advances and governmental regulations relating to cell phones, pagers, ATMs, faxes, computers, Internet access, and local number portability. In order to meet this demand, the North American Numbering Plan (NANP) was modified to allow for what became known as *interchangeable area codes*. Originally an area code could not be the same as a prefix number, and a prefix number could not be the same as an area code number. Starting in 1973, prefix codes could be any number in the format NXX where N is a number between 2 and 9 and X a number between 0 and 9. In 1995, area codes were allowed to be any number in the same format (NXX).

Interchangeable area codes are also allowed for area code overlays, where two or more different area codes can service the same geography. In some areas, the borders for new area codes have been drawn to conform to legal geographic boundaries, such as city or county lines. Instead of moving entire prefixes to a new area code, prefixes were "partitioned" into two or three pieces across two or three different area codes. Today, an exchange can be serviced by prefixes in multiple area codes, in which case 10-digit dialing is mandatory. For this reason, many people refer to the NPA-NXX as the prefix.

Historically, a prefix was assigned to a single service provider. In order to conserve the pool of available telephone numbers, 1000-block pooling was introduced, requiring that unused blocks of local numbers be made available for reassignment to other service providers. A 1000-block is a block of 1,000 consecutive local numbers within a prefix, all starting with the same digit; for example, 203-255-1000 through 203-255-1999. As a result, a single prefix may now contain multiple carriers, providing different types of service to different-sized exchange areas.

Linda Piekarski

See also Access Lines; Number Portability; Suffix Banks

Pre-Primary Polls

Pre-primary polls are those conducted before primary elections in an attempt to measure voters' preferences. They present a number of difficult challenges for pollsters.

Voters in primary elections can be volatile in their preferences, often because many candidates are relatively unknown and voters have limited information about them. This is especially true in early primary states in U.S. presidential campaigns, when the field of candidates is often wide and there are many lesser-known candidates struggling to gain momentum. With little information, voters can be highly susceptible to the dynamics of the campaign, and every time a candidate starts a new ad campaign or launches a microtargeting effort, poll numbers can move in response.

Primary electorates are made up of partisans of one party (or of independents who lean toward a party) and thus tend to be relatively homogeneous as compared to the electorates in general elections. It can be easier to predict the outcomes of general elections because partisans of one party generally choose to vote for their party's candidates, making their votes relatively easy to predict. Pollsters in general elections can focus on studying how independents and swing voters will behave on Election Day. But in primary elections, voters base their decisions on a complex set of preferences that go beyond basic party affiliations and are much harder to quantify.

Primary elections also often have low or variable turnout, making turnout projections difficult. While turnout in heavily contested primaries early in the presidential cycle sometimes can rival turnout in general elections, turnout in primaries is usually much lower. Many primaries are uncontested or are not very competitive, because incumbents are rarely challenged. Thus when a competitive primary comes along, there is often little voting history that can be useful in predicting future turnout. Turnout models for general elections that place importance on past voting behavior often are not as good at predicting turnout in primaries. And when competitive primaries happen in areas that have not seen competition in a while, they often draw large numbers of people who have not voted in a primary before. The behavior of first-time voters can be difficult to predict.

U.S. election laws and party regulations related to the administration of primary elections and party caucuses vary dramatically across states, adding to the complexity of preprimary polling. For example, some states allow only voters registered with a given party to vote in that party's primary. These primaries are called "closed primaries." Other states have "open primaries" that allow any voter to choose the primary in which he or she will vote. This makes it important to include political independents in pre-primary polls in open primary states. The fact that independents can change their minds at the last minute about which party's primary they will choose makes it a challenge to predict the composition of each party's electorate.

The fact that primary electorates are made up of one party's partisans also presents difficulties for sampling. Most pre-primary polls use screening techniques to determine eligibility to vote in a particular primary, and such screening techniques can be expensive to administer. In states with closed primaries and party registration, lists of registered voters can be used to draw samples of party registrants, but the quality of such lists varies dramatically by state. Lists can be out of date or lack contact information for the voters. It can also be challenging to develop appropriate weighting schemes for pre-primary polls, because it is difficult to find good estimates of the demographic profiles of a particular primary electorate that can be used as weighting benchmarks.

Polling in caucus states presents special challenges. For example, in Iowa's Democratic Party caucuses, participants gather to select delegates to the party convention in a multi-stage process, which involves an initial stage intended to narrow the field of candidates to a small number of "viable" candidates. Participants who initially supported one candidate upon entering the caucus may realign and choose another candidate to support after the viable candidates are selected. Thus pre-caucus polls must anticipate the second choice of caucus-goers in predicting the outcome of the caucuses.

Many media organizations conduct pre-primary polling as part of their election coverage, but budgets for pre-primary polls are often lower than those for general election polling. Thus media pre-primary polls are often based on small samples, are conducted with limited frequency and with less sophisticated turnout models, all of which can contribute to their volatility and affect their accuracy. Media organizations often conduct pre-primary polling using national samples, which can be useful for analyzing the dynamics of the campaign across all the states but have little predictive power when it comes to the races in individual states.

On the other hand, the national attention focused on early primary and caucus states like Iowa and New Hampshire means that residents of those states are polled early and often by media organizations and candidates alike. Multiple and often contradictory polls are reported in these early primary states. Pollsters compete with campaigns for the time and attention of voters, who by some accounts are driven crazy by telephone calls and emails. The increasing cacophony of polling and electioneering in early presidential primary states presents new problems for pre-primary polling.

Trevor N. Tompson

See also Election Polls; Horse Race Journalism; Leaning Voter; Likely Voter; Media Polls; Pre-Election Polls

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Primacy Effect

The *primacy effect* is one aspect of a well-known phenomenon called the "serial position effect," which occurs when one is asked to recall information from memory. The other aspect of the serial position effect is the *recency effect*. Psychologists discovered these effects more than a century ago; for example, in the 1890s, Mary Whilton Calkins experimented with these effects while she was a student of William James. Essentially, the serial position effect means that the recall of a list of items is easiest for a few items at the end of the list and for a few items at the beginning of the list. The recall of items in the middle of the list is generally poor. The primacy effect refers to the recall of items at the beginning of the list, while the recency effect refers to the recall of items at the end of the list. If one graphed the number of recalled items as a function of position in the list, one would obtain a U-shaped function.

Suppose that an experiment were performed in which the following list of 24 words was read aloud at the rate of one word per second to a group of persons:

> apple, basketball, cat, couch, potato, book, bus, lamp, pencil, glasses, guitar, truck, photo, rose, apartment, movie, clock, car, dog, bowl, shoe, bicycle, plane, university

The individuals who had listened to the words being read aloud were then asked to write down all the words that they could remember. According to the serial position effect, the recency effect would predict that the terms *university, plane*, and *bicycle* would be easily recalled; in addition, the primacy effect would predict that the terms *apple, basketball*, and *cat* would also be easily recalled. However, items in the middle of the list, such as *guitar*, *movie, photo*, and *rose* would be least likely to be remembered.

The only exception to this principal occurs when an item in the middle of the list is extremely well known. For instance, suppose that one were asked to write down all the presidents of the United States that one could recall. The primacy and recency effects are still in evidence; that is, the current president and a few preceding ones are easily recalled, as are several of the first few presidents of the United States. But in this case, Abraham Lincoln is so famous that although his presidency occurred in the middle of the list, he tends to be very easily recalled. And because he is so well known, the presidents associated with him chronologically also tend to have elevated probabilities of being recalled. In this case, the graph of the recall frequency of presidents in chronological order includes two U-shaped graphs, one following the other, where Lincoln represents the peak in the middle of the list of presidents. This is known as the *von Restorff effect* and was discovered in 1933. It does not matter whether research participants are asked to write down all of the presidents that they can remember (a free recall task) or whether they are asked to write down the presidents in order of their respective terms in office; in both cases, the von Restorff effect occurs.

It is important for survey researchers to know about the serial position effect because it can cause response order effects in closed-ended questions. For example, suppose a survey related to an upcoming presidential election includes an item that lists seven Democratic contenders and asks, *For whom would you vote?* If the question is always asked with the candidates' names in the same order, it is likely that one would obtain a U-shaped distribution of responses. To avoid the serial position effect, one needs to use multiple forms of the item so that candidates' names are rotated.

Carla R. Scanlan

See also Measurement Error; Recency Effect; Response Order Effects

Further Readings

Roediger, H. L., & Crowder, R. G. (1982). A serial position effect in recall of United States presidents. In U. Neisser (Ed.), *Memory observed: Remembering in natural contexts* (pp. 230–237). San Francisco: W. H. Freeman.

Primary Sampling Unit (PSU)

In sample surveys, primary sampling unit (commonly abbreviated as PSU) arises in samples in which population elements are grouped into aggregates and the aggregates become units in sample selection. The aggregates are, due to their intended usage, called "sampling units." *Primary sampling unit* refers to sampling units that are selected in the first (primary) stage of a multi-stage sample ultimately aimed at selecting individual elements.

In selecting a sample, one may choose elements directly; in such a design, the elements are the only sampling units. One may also choose to group the elements into aggregates and choose the aggregates in a first stage of selection and then elements at a later stage of selection. The aggregates and the elements are both sampling units in such a design. For example, if a survey is selecting households as elements, then counties may serve as the primary sampling unit, with blocks and households on those blocks serving as the sampling units in subsequent sampling stages of the survey. Thus, *sampling unit* is a term that combines sample selection with the units used as potential selections at any point in the sampling process. For example, in a systematic selection of elements from a list, the elements are the sampling units. The sampling unit contains only one element. But in order to reduce cost of sample selection and data collection, samples may be drawn in multiple stages. Elements are grouped into aggregates (e.g., households on blocks in counties), and a sample of aggregates (e.g., counties) is selected at the first stage. Later elements (e.g., households on blocks) are selected from elements in selected aggregates (e.g., blocks). The aggregates selected at the first stage (e.g., counties) are called "primary sampling units." In multi-stage samples, another level of aggregation may occur within selected primary sampling units, and the aggregates at the second stage (e.g., blocks) are selected, that is, the second stage units. The procedure of creating and selecting aggregates within previously selected aggregates may proceed through several levels of a hierarchy until, at the last stage, the individual element (e.g., households) is selected. The selected elements are chosen only within already chosen aggregates within the hierarchy.

Primary sampling units are not necessarily formed from a list of elements or individuals. They may be conceptual representations of groups for which lists could be obtained, one from each primary sampling unit. For example, suppose a sample of adults living in a large country with widespread population is to be selected, and no list of adults of sufficient quality (e.g., good coverage, covering nearly all adults) is available. Aggregation of adults on an existing list cannot take place. However, units can be formed that are for all intent aggregates of adults within them. A common procedure, *area sampling*, uses geographic areas at successive levels of a geographic hierarchy as sampling units, with implicit aggregation of adults (or other elements) who usually reside within them.

For example, in the United States, a common area primary sampling unit is the county. Within counties, geographic areas such as towns or townships or other administrative units defined by geographic borders can be identified. Within towns or other administrative units, the geography may be further divided into smaller units, such as city blocks or enumeration areas with boundaries formed by streets or highways, rivers, and other relatively permanent, readily identified features. These blocks or enumeration areas are created by a government statistical agency for the purpose of providing census counts of various types of units within them.

The geographic hierarchy from county to town to block may be used as successive sampling units in a survey aimed at ultimately selecting elements that can be associated uniquely with the final stage area unit.

In the first stage of selection to obtain a sample of adults, a sample of counties is selected. The counties from which sample counties are selected are primary sampling units. The selected counties are sometimes called "primary selections."

One could, in principle, create a list of all adults in selected counties and select adults who usually reside in the selected counties in a second and final stage of selection. However, the cost of creating lists of adults at the county level would be prohibitive.

Additional area units—blocks—become sampling units within selected counties and are second-stage sampling units. The hierarchy of units and selection may continue to households and persons within households. At some point, lists of units may need to be created at a stage of selection to continue the process. For example, households within selected blocks and persons within selected households can be listed manually. In such a design, counties, blocks, households, and adults are all sampling units, even though some units are defined and listed by a government agency while others are defined and listed by the survey organization. Only counties are primary sampling units in the sample.

James M. Lepkowski

See also Area Probability Sample; Elements; Multi-Stage Sample; Unit

Further Readings

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Priming

Priming is a psychological process in which exposure to a stimulus activates a concept in memory that is then given increased weight in subsequent judgment tasks. Priming works by making the activated concept accessible so that it can be readily used in evaluating related objects. For example, hearing news about the economy may prime individuals to focus on economic considerations when assessing a president's performance because economic concepts are activated, accessible, and presumably relevant for this type of evaluation. In this way, priming affects the opinions that individuals express, not by changing their attitudes, but by causing them to alter the criteria they use to evaluate the object in question.

Priming is a widely used concept with applications in the fields of psychology, political science, and communication. It is also relevant to survey researchers in that priming can inadvertently occur within questionnaires and interviewing, and surveys can be used to study the priming process.

Survey researchers recognize that their instruments may be susceptible to producing unintended priming effects that could bias key measurements. Inadvertent priming can occur when information presented in one part of the survey activates ideas that are then given increased weight in answering subsequent questions. Research in the Cognitive Aspects of Survey Methodology (CASM) movement suggests that respondents may use cues found within a survey as a way of addressing the cognitive challenges of survey participation. This can happen because respondents often do not have well-formed opinions on many survey topics and therefore have to contemplate a mixture of thoughts that come to mind when a question is posed. Ideally, respondents would search their memories for relevant information that could be used to generate a precise summary judgment. However, respondents may choose instead to reduce their cognitive effort by answering with whatever seemingly relevant information is immediately accessible, including information that may have been primed in earlier parts of the survey. By satisficing in this way, respondents can efficiently generate a serviceable, if not necessarily accurate, response by using little more than the ideas they have recently encountered.

To reduce the potential for inadvertent priming within a survey, researchers often carefully consider decisions about question order, wording, and format. A common strategy is to ask open-ended questions before related closed-ended questions so that the open-ended response is not a mere reflection of the ideas primed by the closed-ended form of the question. For example, political surveys typically ask respondents to assess "the biggest problem facing the nation today" before posing more specific questions about particular policies or political events. This ensures that the initial open-ended response is a measure of perceived major problems that does not inadvertently reflect recently primed considerations. Researchers also may take steps to ensure that the phrasing of their questions does not prime thoughts that will bias responses. They may even proactively seek to reduce priming effects by explicitly asking respondents to consider a number of factors before answering a question.

Surveys have also proven valuable in the study of priming, not only as measurement tools but as means to further understand the priming process. In fact, many survey experiments have used controlled manipulations of question order, wording, and format to measure the strength and effectiveness of priming in various situations. Other studies have reorganized surveys to test past findings that may have been biased by inadvertent priming within the survey. Indeed, survey research has been critical in measuring priming effects and illuminating our understanding of the priming process.

Michael Parkin

See also Closed-Ended Question; Cognitive Aspects of Survey Methodology (CASM); Open-Ended Question; Question Order; Saliency; Satisficing

Further Readings

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A., & Kinder, D. R. (1990). Altering the foundations of support for the president through priming. *The American Political Science Review*, *84*, 497–512.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

PRIOR RESTRAINT

Prior restraint refers to a legal principle embodied in the First Amendment to the U.S. Constitution that relates to guarantees of freedom of the press. At the most fundamental level, it provides protection against censorship by the government, and it is particularly relevant to survey research because of legal disputes about the presentation of exit poll results on Election Night.

Prior restraint actually refers to any injunction that would prohibit the publication of information, an infringement on the so-called freedom of the press. Many consider it an especially serious issue because it prevents the dissemination of information to the public, as distinct from an injunction issued after the information has been released that prohibits further dissemination or provides for some kind of relief as in the instance of libel, slander, or defamation of character.

The case law history of the principle is generally credited as starting with *Near v. Minnesota*, in which the U.S. Supreme Court decided in 1931 that a small Minnesota paper could not be prevented in advance from publishing information about elected officials. In this 5–4 decision, the Justices left open the possibility of prior restraint in some cases, especially those involving national security. Censorship was practiced during World War II, but the issue of national security was not addressed explicitly until *The New York*

Times Co. v. United States, a case in which the Nixon administration went to the court in 1971 to prohibit the publication of the Pentagon Papers by *The Washington Post* and *The New York Times*. The Court decided that an injunction was not warranted.

Another important area of litigation involved censorship of artistic content in the theater and films, usually related to obscenity. In order to avoid new legislation and formal constraints, many industry groups developed so-called voluntary codes that circumscribed content. This approach was adopted in the film and comic book industries, and a rating system developed by the Motion Picture Association of America is still in force. Many content producers do not see any significant difference between a formal law and the prior restraint that results from such practices.

It is in this context that the states and the Congress have unsuccessfully attempted to limit or prohibit the collection and dissemination of exit poll data since 1980. This contentious issue arose from a number of considerations. They include the geopolitics of American presidential elections, the role of the Electoral College, and the fact that the election takes place across four time zones in the continental United States and six time zones if Alaska and Hawaii are included. The exit poll data are used to project the outcome of the presidential contest in each state, since the Electoral College generally involves a "winner takes all" allocation of a state's electoral votes, and then each state projection is added to a running total of each candidate's electoral votes. In recent presidential elections, the Democrats have done well on the east and west coasts and contested the Great Lakes region. The Republicans have done well in the South, Midwest, and West, not including the west coast, and they have contested the Great Lakes too.

This geography is obviously closely aligned with the time zones in the United States, and if the Democratic candidate has not built a sizable electoral vote lead by 9 or 10 p.m., they generally cannot achieve a majority. And herein lies the problem, as the television networks can call one state at a time as their polls close and then eventually a projected winner before the voting is completed on the west coast and beyond. In the 1980 presidential election, for example, Ronald Reagan was projected as the winner by NBC at 8:15 EST (5:15 PST), and Jimmy Carter conceded the election an hour and a half later, more than one hour before the polls closed on the west coast. While the Democrats cried "foul" because they believed that turnout on the west coast was depressed because of the early call, Carter's concession after all three major networks came to the same conclusion took them off the hook somewhat. But in the 2000 election, it was the Republicans' turn to claim injury after Florida was called relatively early for Al Gore, and then the call was reversed, and then reversed again. Florida actually votes in two time zones, so the claim was made that turnout in the western panhandle of the state dropped off in the last hour of voting.

There is variety of polling data that show that Americans are opposed to the use of exit polls as part of Election Night coverage and would support restrictions on their use, including banning them altogether. Starting after the 1980 election, buoyed by their own partisan concerns about their accuracy and impact, members of Congress and various state legislatures began to consider legislation to either prohibit exit polling outright or to limit the mode of data collection in a way that made it less feasible and more prone to error. These state and local laws required exit poll interviewers to stay far enough away from a polling place so as to make the systematic sampling of voters leaving the precinct impractical, because voters could not be intercepted efficiently before they got to their vehicles and left the area.

The main defense of the networks was their First Amendment right to gather and disseminate news and the case law opposed to prior restraint except under the direst circumstances. Several bills were introduced in Congress to eliminate exit polls or to restrict the dissemination of news based upon data collected through them. None of these laws passed, mainly because of the clear understanding of their unconstitutionality under the prior restraint doctrine. As a second tack, Congress also considered federal changes to election law to stipulate a standard closing hour for the polls across the nation, a proposal supported by the networks. However, Congress does not have much authority to regulate elections because they are generally considered a state matter. And there is also considerable disagreement over what would be the best set of standardized hours to hold elections.

Some states passed laws that would have kept exit poll interviewers at least 300 feet from a polling place, but this was declared unconstitutional as a constraint on collecting information about political views and disseminating them to citizens, a form of protected speech. This principle has been endorsed repeatedly in a series of state court decisions. On the other hand, a House Concurrent Resolution passed in 1985 that asked the networks to refrain from projecting the outcome in any single race until the polls were closed. This proposal was not very different from a pledge that network executives had already made not to call a state until all or most of the polls in a state were closed. But in effect, an agreement for self-restraint proved to be an acceptable solution to a problem in which Congress and the states were essentially powerless to intervene because the networks were protected by the principle of prior restraint.

Michael W. Traugott

See also Election Night Projections; Election Polls; Exit Polls

Further Readings

Prior Restraints. (2003). *The First Amendment handbook.* Arlington VA: Reporters Committee for Freedom of the Press. Retrieved December 29, 2006, from http://www.rcfp.org/handbook/c05p01.html

Seager, S. E., & Handman, L. R. (2001). *Congress, the networks, and exit polls.* American Bar Association, Forum on Communication Law. Retrieved December 29, 2006, from http://www.abanet.org/forums/communication/comlawyer/winter01/seager.html

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

PRIVACY

Within the realm of survey research, privacy consists of the right to control access to one's self and one's personal information. Private behavior occurs in a context in which an individual can reasonably expect that no observation or recording is taking place. Privacy is distinct from *confidentiality* in that privacy refers to the protection of the right of individuals, whereas confidentiality refers to the protection of the data collected.

Identity, health and financial information, criminal justice involvement and court records, education, and work performance data are commonly regarded as private, despite the fact that many are commonly accessible through credit-reporting background check agencies.

The distinction between public and private behaviors is often ambiguous. Some information that becomes part of public record, such as a person's phone number, marital status, or income, would be considered private information and in the United States is covered by a federal law, the Family Educational Rights and Privacy Act, also known as FERPA or the Buckley Amendment. Some behaviors that occur in the public realm, such as a person's presence in a "red-light district," hospital emergency room, stock brokerage, or welfare office may also be considered private.

An individual's ability to control access to their personal information is often determined by socioeconomic status, age, and other circumstances. For example, information about receiving welfare or credit history is public, whereas account balances are not, unless you are a government official. Minors have generally fewer rights to privacy than do adults. Institutionalized persons may have significant limitations on their ability to control personal information.

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule regulates the disclosure of information about health status, provision of health care, or payment for health care that can be linked to an individual. Student education records are protected by the FERPA. In the United States, however, there is no comprehensive federal law protecting all the private records and information of individuals. In contrast, the European Directive on Data Protection requires countries in the European Union (EU) to have government data protection agencies, requires registration of databases with those agencies, and in some cases requires prior approval before personal data processing begins. In order to bridge these different privacy approaches and provide a streamlined means for U.S. organizations to comply with the directive, the U.S. Department of Commerce in consultation with the European Commission developed a "safe harbor" framework. The safe harbor—approved by the EU in July 2000—provides direction for U.S. companies to help them comply with European privacy laws.

Institutional review boards (IRBs) provide critical oversight for human subjects research to ensure that research does not constitute an invasion of privacy. An individual's right to privacy from research, including survey research, is generally protected by the right to refuse to participate in research or to refuse to answer any individual question with a survey or interview. Controversy issues arise when investigators wish to use personally identifiable records or observe behavior without obtaining consent. In general, if data are publicly available and cannot be linked to any identifiable subjects, there are no privacy concerns.

Federal regulations require IRB review of government-funded research studies that include observations of public behavior that could identify individual subjects and, if made public, could reasonably place the subject at risk of criminal or civil liability or cause damage to the subject's financial standing, employability, or reputation. The IRB will generally weigh these risks against the value of the knowledge to be gained.

Once private data are collected, secure storage becomes an important obligation of the researcher. In accordance with HIPAA requirements, private data should be protected by storing backups in a separate location, securing computers (both workstations and servers) and storage devices with locks, protecting computers and electronic media with "sign-on" passwords, and by using encryption software to encode private data.

Amy Flowers

See also Certificate of Confidentiality; Confidentiality; Informed Consent; Institutional Review Board (IRB); Perturbation Methods; Voluntary Participation

Privacy Manager

The privacy manager sample disposition is specific to telephone surveys. Privacy manager services are telecommunications technologies that are available to households and businesses as an optional subscription from most local U.S. telephone companies. Many privacy manager services work with caller identification (caller ID) services to identify and block incoming calls that do not display a telephone number that the household has identified as a known number. Because very few households or businesses will have the name or telephone number of a telephone survey interviewing organization as a "known" number, privacy manager technologies make it significantly more difficult for telephone interviewers to contact households that subscribe to these services.

In addition to blocking calls from "unknown" numbers, most privacy manager services also block calls whose identification is displayed as anonymous, unavailable, out of area, or private. Callers with blocked numbers usually have the option to temporarily unblock their numbers so their calls can be connected. When interviewers dial a number that has privacy manager services, the interviewer will hear something like, "The person you are trying to reach does not accept unidentified calls. Your caller ID/telephone number was not received/known." Interviewers then have the opportunity to identify who is calling. If the interviewer announces his or her name and organization, usually the phone then will ring and the household or business will hear the identification provided by the interviewer. At that point, the household or business can choose to accept the call, reject the call, send the call to a voice-mail system or answering machine, or send a "solicitor's rejection," such as notifying the caller that "telemarketing calls are not accepted."

The number of households and businesses with privacy manager services has increased as the number of telemarketing calls has grown in the past decade and as the cost of these services has dropped. As a result, many survey organizations have established "privacy manager" as a unique survey disposition—both to track the prevalence of these call outcomes and to help ensure that cases with these outcomes are managed properly. Finally, because privacy manager technologies make it more difficult to screen numbers in a sample, cases that have a call outcome of privacy manager usually are considered cases of unknown eligibility (because many times it is difficult or impossible for telephone interviewers even to determine whether the case is a household or not). Existing evidence suggests that the level of cooperation that eventually can be gained from repeated calling attempts to households with privacy manager is very low ($<5\%$).

Matthew Courser

See also Caller ID; Dispositions; Final Dispositions; Nonresidential; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Probability

In general, probability is a numerical representation of how likely is the occurrence of certain observations.

Whenever an empirical investigation involves uncertainty, due to sampling, insufficient understanding of the actual procedure or of the laws governing the observed phenomena, or for any other reason, the concept of probability may be applied.

The concept of probability developed from the investigation of the properties of various games, like rolling dice, but in addition to the very practical desire of understanding how to win, it also incorporates deep philosophical thought. Currently, most scholars consider probabilities as objectively existing values that can be best revealed by observing long sequences of potential occurrences of events and using the relative frequencies of the events to approximate their probabilities. Another, also objective, view of probability is that it can be calculated as the ratio of the possible number of observations when an event occurs to the total number of possible observations. Some other scholars think that probability is subjective: It expresses the degree of belief a certain person has in the occurrence of an event. Fortunately, all these approaches lead to probabilities that have essentially the same properties.

A simple illustration is rolling a fair die with all outcomes being equally likely. Then, each outcome has probability equal to 1/6. More precisely, this may be a subjective belief of an observer, or the assumption of the experimenter, or may be empirically tested by observing the results of long sequences of rolls. In either case, the probability of having a value less than 4 (i.e., 1, 2, or 3) is equal to ½, and the probability of having an even outcome (i.e., 2, 4, or 6) is also ½. The probability of having an outcome less than 4 *and* even is 1/6, as this only happens for 2. There are however, events that are possible but cannot occur at the same time. For example, having a value less than 4 and greater than 5 is not possible. The event that cannot occur is called the "impossible event" and is denoted by Ø.

Probability theory has many applications in the social sciences. It is the basis of random sampling, where the units of the population of interest, usually people, are selected into the sample with probabilities specified in advance. The simplest case is simple random sampling, where every person has the same chance of being selected into the sample and the steps of the selection process are independent from each other. In such cases, probability models with all outcomes having equal probability may be relevant, but the more complex sampling schemes often used in practice require other models. Probabilistic methods are also used to model the effects of errors of a measurement and, more generally, to model the effects of not measured or unknown factors.

Theory of Probability

In a precise theory, the events associated with the observation of an experiment possess probabilities. To define the properties of probability, one has to define certain operations on these events. The product of two events A and B occurs if both A and B occur and is denoted by AB. For example, one event may be, when rolling a die, having a value less than 4; another event may be having an even value. The product of these is having a value that is less than 4 and is even, that is, 2. Or, an event may be hitting the right-hand side of a circular target, another may be hitting its upper half, and the product is hitting the upper right-hand-side quarter. If two events cannot occur at the same time, their product is the impossible event, Ø. Another operation is the sum of two events, denoted as A + B. The sum occurs if at least one of the two original events occurs. The sum of having a value less than 4 and of having an even value is having anything from among 1, 2, 3, 4, and 6. The sum of the right-hand-side of the target and of its upper half is three quarters of the target, obtained by omitting the lower left-hand-side quarter.

There are three fundamental rules governing probabilities:

1. The probability of an event is a number between 0 and 1:
$$0 \le P(A) \le 1.$$
2. If an event occurs all the times (surely), its probability is 1:
$$P(S) = 1.$$
3. If two events are such that they cannot occur at the same time, then the probability of their sum is the sum of their probabilities:
$$P(A + B) = P(A) + P(B), \text{ if } AB = \emptyset.$$

The last property may be extended to cover the sum of infinitely many events. To see that this may be necessary, think of how many times a die needs to be rolled to have the first 6. Let A1 be that once, A2 that twice, and so on. One cannot say that at most A10 or A1000 cannot occur. The event that one has a 6 is A1 + A2 + A3 + ⋯ A more general form of 3 postulates:

3. If a sequence of events A1, A2, A3,... is such that no two events can occur at the same time, then the probability of their sum is the sum of their probabilities, that is, $P(A1 + A2 + A3 + \ldots) = P(A1) + P(A2) + P(A3) + \ldots$ if $AiAj = \emptyset$ for $i \neq j$.

Independence

The events A and B are independent, if

$$P(AB) = P(A)P(B).$$

In the die-rolling example, the events of having an even number and of having a value less than 5 are independent, because the former (2,4,6) has probability 1/2 and the latter (1,2,3,4) has probability 2/3, while the intersection (2,4) has probability 1/3, that is, the product of 1/2 and 2/3. On the other hand, the events of having an even number and having something less than 4 are not independent, because the former has probability 1/2, the latter has probability 1/2, and their product has probability 1/6. The interpretation of this fact may be that within the first three numbers, even numbers are less likely than among the first six numbers. Indeed, among the first six numbers, three are even numbers; among the first three, only one is even.

Conditional Probability

The conditional probability shows how the probability of an event changes if one knows that another event occurred. For example, when rolling the die, the probability of having an even number is 1/2, because one half of all possible outcomes are even. If, however, one knows that the event of having a value less than 4 has occurred, then, knowing this, the probability of having an even number is different, and this new probability is called the "conditional probability" of having an even number, given that the number is less than 4. Because there is only one even number out of the first three, this conditional probability is 1/3.

The conditional probability of A given B is denoted by P(A | B) and is precisely defined as

$$P(A|B) = P(AB)/P(B),$$

that is, the probability of the product of the events, divided by the probability of the condition. We have seen that the probability of having a value that is even and also less than 4 is 1/6, and the probability of having a value less than 4 is 1/2, and their ratio is 1/3.

It follows from the definition directly that if A and B are independent, then P(A | B) = P(A), that is, conditioning on the occurrence of B does not change the probability of A, if these are independent. Indeed, if A and B are independent,

$$P(A|B) = P(AB)/P(B) = P(A)P(B)/P(B) = P(A),$$

where the first equality is the definition of conditional probability and the second one is based on the definition of independence. Therefore, independence of events means no relevance for each other.

Applications

The foregoing simple examples are not meant to suggest that formulating and applying probabilistic models is always straightforward. Sometimes even relatively simple models lead to technically complex analyses. Further, the inference based on a probabilistic model is only relevant if, in addition to the correct analysis of the model, it is also based on a model that appropriately describes the relevant aspects of reality. The results will largely depend on the choice of the model.

Tamás Rudas

See also Frequency Distribution; Percentage Frequency Distribution; *p*-Value; Random Error; Random Sampling; Relative Frequency; Simple Random Sample

Further Readings

Bennett, D. J. (1998). *Randomness.* Cambridge, MA: Harvard University Press.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). New York: Wiley.

Rudas, T. (2004). *Probability theory: A primer*. Thousand Oaks, CA: Sage.

Rudas, T. (Ed.). (2007). *Handbook of probability theory with applications.* Thousand Oaks, CA: Sage.

Probability of Selection

In survey sampling, the term *probability of selection* refers to the chance (i.e., the probability from 0 to 1) that a member (element) of a population can be chosen for a given survey. When a researcher is using a probability sample, the term also means that every member of the sampling frame that is used to represent the population has a known nonzero chance of being selected. That chance can be calculated as a member's probability of being chosen out of all the members in the population. For example, a chance of 1 out of 1,000 is a probability of 0.001 $(1/1{,}000 = 0.001)$. Since every member in a probability sample has some chance of being selected, the calculated probability is always greater than zero. Because every member has a known chance of being selected, it is possible to compute representative unbiased estimates of whatever a researcher is measuring with the sample. Researchers are able to assume with some degree of confidence that whatever they are estimating represents that same parameter in the larger population from which they drew the sample. For nonprobability samples (such as quota samples, intercept samples, snowball samples, or convenience samples), it is not feasible to confidently assess the reliability of survey estimates, since the selection probability of the sample members is unknown.

In order to select a sample, researchers generally start with a list of elements, such as addresses or telephone numbers. This defined list is called the "sampling frame." It is created in advance as a means to select the sample to be used in the survey. The goal in building the sampling frame is to have it be as inclusive as possible of the larger (target) population that it covers. As a practical reality, sample frames can suffer from some degree of undercoverage and may be plagued with duplication. Undercoverage leads to possible coverage error, whereas duplication leads to unequal probabilities of selection because some elements have more than one chance of being selected. Minimizing and even eliminating duplication may be possible, but undercoverage may not be a solvable problem, in part because of the cost of the potential solution(s).

In designing a method for sampling, the selection probability does not necessarily have to be the same (i.e., equal) for each element of the sample as it would be in a simple random sample. Some survey designs purposely oversample members from certain subclasses of the population to have enough cases to compute more reliable estimates for those subclasses. In this case, the subclass members have higher selection probabilities by design; however, what is necessary in a probability sample is that the selection probability is knowable.

Depending on the method of data collection, the final selection probability may not be known at the outset of data collection. For example, in household surveys, such as those selected via random-digit dialing (RDD), additional information such as the number of eligible household members needs to be collected at the time of contact in order to accurately compute the final selection probability. The more eligible members in the household, the lower is the selection probability of any one member; for example, in a household with a wife, husband, and two adult children, each has a probability of selection within their household of 1/4. Furthermore, in RDD landline telephone surveys of the general public, it is common to ask how many working telephone numbers are associated with a household. If there are two working landline telephone numbers, then the household has twice the chances of being selected compared to households with only one working landline number, and thus a weighting adjustment can be made for households with two or more numbers. Similarly, in mail survey questionnaires that are not sampling specifically named people, a question about household size regarding eligible members is generally asked. In a systematic sample (e.g., exit polls), the probability of selection is the inverse of the sampling interval.

Selection weights, generally the inverse of the selection probability for each case, are calculated to adjust for differential selection probabilities due to features of the sample design (e.g., stratification) or sampling strategies like oversampling subclasses. To compensate for most complex design features, special statistical software solutions need to be employed. In a similar fashion, using the inverse of the selection probabilities, response rates should also be weighted when unequal selection probabilities exist.

Mario Callegaro and Charles DiSogra

See also Complex Sample Surveys; Convenience Sampling; Coverage Error; Duplication; Element; Mall Intercept Survey; Nonprobability Sample; Overcoverage; Post-Stratification; Probability; Probability Sample; Quota Sampling; Response Rates; Sampling; Sampling Frame; Sampling Interval; Simple Random Sample; Snowball Sampling; Systematic Sampling; Target Population; Undercoverage; Weighting

Further Readings

Levy, P. S., & Lemeshow, S. (2008). *Sampling of populations: Methods and applications* (4th ed.). Hoboken, NJ: Wiley.

Probability Proportional to Size (PPS) Sampling

Probability proportional to size (PPS) sampling includes a number of sample selection methods in which the probability of selection for a sampling unit is directly proportional to a size measure, X_i, which is known for all sampling units and thought to be approximately proportional to the unknown Y_i. The X_i must all be strictly greater than 0. In single-stage sampling it can be used to reduce the variance of survey estimates. If the observed values, Y_i, are exactly proportional to X_i , the variance of an estimated total will be exactly 0. When the Y_i are approximately proportional to the X_i, the variance can still be greatly reduced relative to equal probability sampling schemes. PPS sampling is also used in the early stages of multi-stage samples to achieve equal probability samples of the final-stage sampling units or EPSEM samples.

In all cases, suppose one is estimating a population total for a variable Y_i with N units.

$$Y = \sum_{i=1}^{N} Y_i.$$

There are a wide variety of sampling schemes that have PPS properties. Only a few of them are discussed here for illustration.

PPS With Replacement Sampling

The simplest PPS sampling method is *PPS with replacement* sampling. In this case, the single-draw probability of selection, p_i, on each independent sample draw is proportional to X_i, that is,

$$p_i = \frac{X_i}{\sum_{i=1}^{N} X_i}.$$

Note that unless $N = 1$, the individual probabilities will be less than 1. The with replacement estimator is

$$\hat{Y}_{wr} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i},$$

where y_i represent the observed values indexed over the sample of n draws. Note that the same population unit may be selected at more than one draw. The variance of this estimator is

$$V(\hat{Y}_{wr}) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{Y_i}{p_i} - Y\right)^2.$$

The replacement variance estimator has a simpler form:

$$v(\hat{Y}_{wr}) = \frac{1}{n(n-1)}\sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{wr}\right)^2.$$

An unbiased estimate of the variance can be obtained for all samples of size 2 draws or greater.

To select a PPS with replacement sample, the following procedure is followed for each sample draw. Select a uniform (0,1] random number, R. Assume the N population elements are ordered and indexed by I. Compute the partial sums $S_i = \sum_{j=1}^{i} p_j$. Select the unit i if $S_{i-1} < r \le s_i$; in simpler terms, the unit selected will be the first one whose partial sum equals or exceeds R.

PPS Without Replacement Sampling

In practice, *PPS without replacement* sampling is more commonly used, particularly in the first stage of multi-stage samples. Since sampling is without replacement, the sample size n now represents unique population units selected into the sample. Let S be the set of all $\binom{N}{n}$ samples of size n from the population of size N and is indexed by s and $\sum_{s \in S} P_s = 1$. P_s represents the probability of a particular sample s. The probability of selecting any particular unit i in a sample of size n is formally defined as $\pi_i = \sum_{s \supset i} P_s$. Summation is over all samples of size n that contain the population element i. Many creative methods have been developed to select PPS without replacement samples that satisfy the property $\pi_i = np_i$ where p_i is as defined previously for with replacement sampling. It often happens that $np_i \ge 1$ for some i. A common fix for this problem is to select unit

i with probability 1 and select a PPS without replacement sample of size $n-1$ from the remaining elements.

The commonly used estimator associated with PPS without replacement sampling is the *Horvitz-Thompson estimator*, $\hat{Y}_{wpr} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$.

In order to compute the variance of the Horvitz-Thompson estimator, it is also necessary to know the pairwise selection probabilities. These probabilities, denoted by π_{ij}, are the probabilities that population units i and j both get selected in the same sample. Similarly to the formal definition for the unit probabilities, the pairwise selection probabilities are defined in terms of the sum over a subset of sample selection probabilities; in this case, the summation is limited to samples that contain both unit i and unit j:

$$\pi_{ij} = \sum_{s \supset (i,j)} P_s.$$

The Yates-Grundy form of the variance of the Horvitz-Thompson estimator is

$$V(\hat{Y}_{wor}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j \neq i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2.$$

This form of the variance is only defined for samples of size 2 or greater. Samples of size 1 meet the PPS with replacement definition.

The variance estimator has similar form:

$$v(\hat{Y}_{wor}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j \neq i}^{n}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Unbiased estimates of the variance can be obtained only if the pairwise probabilities, π_{ij}, are positive for all pairs of units.

PPS Methods for Small Samples

A number of sampling schemes have been developed for selecting PPS without replacement samples that have all positive pairwise probabilities. Since PPS sampling is often used in conjunction with stratification, samples of size 2 per stratum are not uncommon. For samples of size 2, simply select the pair with probability

$$\pi_{ij} = K\pi_i\pi_j\left[\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j}\right] \text{ where}$$

$$K = \left[2 + \sum_{i=1}^{N}\frac{1}{1-\pi_i}\right]^{-1}.$$

M. R. Sampford's rejective method may be applied to larger samples and also produces positive pairwise probabilities. In this method of selection, an initial draw is made with PPS exactly as in the method used for each draw in PPS with replacement sampling with probabilities p_i. The remaining $n-1$ selections are also drawn using PPS with replacement with probabilities proportional to $\lambda_i = \frac{p_i}{1-np_i}$. If the first n draws yield unique elements, then one retains the sample. Otherwise, one repeats the entire process until a sample of n unique elements is selected.

PPS Systematic

PPS systematic sampling provides a relatively simple method for selecting larger PPS samples but does not provide for unbiased variance estimation since most pairwise probabilities will be 0. For larger samples, the ordering of the frame before selecting the sample is often desirable, since it imposes an implicit stratification along the ordering dimension. For example, some socioeconomic balance can be achieved in a sample by ordering counties or other geographic units based on percentage above or below a specified income level in the most recent decennial census. Approximate variance formulas that recognize the implicit stratification are often used.

PPS sample selection is relatively simple. Note that

$$\pi_i = \frac{nX_i}{\sum_{i=1}^{N} X_i},$$

and the π_i sums to n. To select a PPS systematic sample from an ordered list, select a uniform (0,1] random number, R. Then form partial sums of the ordered list $S_i = \sum_{j=1}^{i} \pi_j$. Select the n units i that satisfy $S_{i-1} < r + k - 1 \leq s_i$ for $k = 1, 2, \ldots, n$.

Of note, PPS systematic sampling can be viewed as a sequential sampling method.

James R. Chromy

See also EPSEM Sample; Equal Probability of Selection; Multi-Stage Sample; Sampling Without Replacement;

Sequential Sampling; Stratified Sample; Systematic Sampling; Unit

Further Readings

Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities.* New York: Springer.
Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory.* New York: Wiley.
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.
Sampford, M. R. (1969). A comparison of some possible methods of sampling from smallish populations with units of unequal size. In N. L. Johnson & H. Smith, Jr. (Eds.), *New developments in survey sampling*. New York: Wiley-Interscience.

Probability Sample

In gathering data about a group of individuals or items, rather than conducting a full census, very often a sample is taken from a larger population in order to save time and resources. These samples can be classified into two major groups describing the way in which they were chosen: *probability samples* and *nonprobability samples*. Both types of samples are made up of a basic unit called an individual, observation, or elementary unit. These are the units whose characteristics are to be measured from a population. In probability samples, each member of the population has a known nonzero probability of being chosen into the sample. By a random process, elements are selected and receive a known probability of being included in the sample; this is not the case in nonprobability sampling.

In order to estimate some quantity of interest with a desired precision from a sample, or to contrast characteristics between groups from a sample, one must rely on knowing to whom or to what population one is referring. Well-designed probability samples ensure that the reference population is known and that selection bias is minimized. The best samples are simply smaller versions of the larger population. The process by which a sample of individuals or items is identified will affect the reliability, validity, and ultimately the accuracy of the estimates and inferences made.

Underlying Concepts

The concepts behind probability sampling underlie statistical theory. From the finite population of N elements, all possible samples of size n are identified. For example, if the population consisted of 6 elements, and samples of size 2 are to be chosen, there would be 15, $\binom{6}{2}$, possible samples to consider. In theory, prior to selection, the probability of each of these samples being chosen is known. Therefore, the selection probability of each individual is also known. Summing the probabilities of all samples containing an individual element will compute the individual's probability of appearing in the sample. By knowing the probability of selecting each unit, a statistical weight can be assigned by which population estimates can be calculated. The statistical weight is defined as the inverse of the probability of selection into the sample, allowing each sampled unit to represent a certain number of units in the population. Most often, the goal of a probability sample is to estimate certain quantities in the population using these statistical weights.

Reliability and Validity

From a probability sample, the quantities that are estimated, such as population totals, means, proportions, and variances, have certain properties that can be evaluated. For instance, over repeated sampling, estimators from the probability sample can be evaluated for how reproducible or reliable they are (variance) and if, on average, the estimates from the sample are similar to the true value of the population quantity (unbiasedness or validity). Combining the ideas of reliability and validity, the accuracy, or how far away on average the estimator is from the true value (mean squared error), can also be evaluated on the sample estimator.

None of these desirable properties can be determined from estimates derived from nonprobability samples. Nonprobability samples are used in many unscientific surveys, market research, and public opinion polls, often because they are easier and less expensive to conduct. These types of surveys include purposive or deliberate, quota, and snowball samples. As an example, imagine that interviewers are attempting to question shoppers as to their political views at a local supermarket in order to describe future poll results for a city. In quota sampling, the interviewer may be asked to "find" a certain number of individuals in various demographic groups, such young women, older women, black men, and white men. The individuals that are found by the interviewer may be of only one political leaning or of one socioeconomic group simply because they are easy to find and shop at the particular market. Without a systematic plan for sampling, the decision about whom to interview is left up to the interviewer, likely creating bias in the sample. In certain circumstances, such as in populations that are hard to reach, probability samples may not be feasible. Thus, as long as they are not used to make inference to a larger population, some nonprobability samples are useful.

In order to draw a probability sample, a list of sampling units must be organized from which to sample. In hard-to-reach populations, this sampling frame may not be available. The sampling frame must allow every unit in the population to have a chance of selection; otherwise coverage error could result. For example, the sampling frame may consist of all medical records within a certain time period. A sample can be drawn from the medical record numbers. Alternatively, the sampling frame may be an enumerated list of city blocks from which a household, and eventually an individual, will be chosen. Using the sampling frame, the sample can be selected in numerous ways.

Sampling Designs

Simple Random Sampling

The simplest design used to sample units is called a "simple random sample." Simple random sampling can either be done *with replacement* or *without replacement*. In sampling with replacement, the unit sampled can be selected more than once, since it is returned to the pool once sampled. In practice, nearly all probability samples are performed without replacement, having each element appear only once in the sample. Simple random samples without replacement consist of selecting n units from a population of N elements, each possible subset having the same probability of selection. For example, if numbers 1 to 100 were placed in an urn, and 10 numbers were drawn without replacing them after each turn, this would be a simple random sample. The units or people associated with these 10 numbers would constitute the sample from the population of size 100.

In fact, simple random samples are not often used alone as the sample design for a survey. Enumerating every element of a population is a tedious and time-consuming process. In addition, once this list is compiled it may be out of date. Perhaps a study's interest is in estimating a certain health characteristic of high school students in a city. If one were to enumerate all high school students in the city today, students may leave or enter the school tomorrow. Moreover, performing a simple random sample design for a population that covers a large geographic area is not practical. Since each element will have the sample probability of selection, it may require traveling many miles between elements. For all of these reasons, simple random sampling is rarely used alone; however, it provides the basis for comparison to other more commonly used designs.

Systematic Sampling

In some situations, it is not possible to construct a sampling frame before sampling has to occur, but a very good estimate of the number of records in a certain interval may be known. In situations like these it is possible to take a probability sample comprised of every kth element in the population. This is called "systematic sampling." Suppose characteristics of patients presenting to an emergency room is being planned and, from previous observation, it is known that 1,000 patients will come to the emergency room in a given week. If we would like to collect information on 100 patients using a systematic sample, we will survey every 10th patient. A random number is selected between 1 and 10, and every 10th element after that is included in the sample. Systematic sampling is an extremely popular sampling method due to its ease of implementation.

Stratified Random Sampling

One way to improve the precision of estimates over what is possible with simple random sampling is to carry out what is called a "stratified random sample." In this type of sample, elements of a population may also be categorized into distinct groups, called "strata." Within each stratum, a sample of units can then be drawn using simple random or systematic sampling. In general, stratification by a characteristic will reduce variability in the resulting population estimates, especially when the characteristic is related to the measurement of interest. Often, it will also allow reliable estimates to be made about each stratum. Following from the previous example, the individuals to sample may be either male or female. Separately, a simple random sample of men and a simple random sample of women can be chosen, stratifying the sample by gender. Disadvantages of stratified random sampling still exist, including the cost of constructing a sampling frame for each stratum before drawing the sample and then a probability sample from each. This may actually increase the costs of selecting the sample over what would be the case for a simple random sample, but the increased precision may justify the additional time and cost.

Cluster Sampling

One common variation in designing a probability sample is called "cluster sampling." In this type of sampling, the sampling frame enumerates listing units that are not the individuals or elements, but are larger groupings called "clusters." For instance, it may be easier to enumerate city blocks and sample these blocks rather than enumerate all individuals to conduct a survey describing the health of individuals in an area. Once blocks are chosen, dwelling units can be sampled and individuals selected within each unit. Cluster sampling may dramatically reduce cost and time; however, these designs usually have a trade-off in precision. As opposed to simple random sampling, sampling frames are much easier to construct for cluster samples. Only the elementary units of the chosen clusters will have to be listed for sampling. This may entail counting the number of houses on a city block or students in a particular classroom as opposed to the entire city or school. Although clusters may be chosen randomly, most often they are chosen with probability proportional to some measure of the cluster's size (PPS). In general, cluster sampling with PPS reduces variability in estimates as compared to cluster sampling with equal probability, when the size of clusters varies greatly. Moreover, since cluster samples greatly decrease travel costs, the decreased precision for a fixed sample size compared to simple random sampling is outweighed by the ability to sample a greater number of individuals for a fixed cost, ultimately resulting in more precise estimates for a fixed budget. Cluster sampling is feasible in many situations where simple random sampling is not.

Some common *random-digit dialing* (RDD) techniques are, in fact, cluster samples. The most straightforward random-digit dialing method simply randomly selects phone numbers from a frame: a simple random sample. However, only 20% of the phone numbers may be household phone numbers. RDD methods based on sampling combinations of the area code, the exchange, and blocks of the remaining four numbers are cluster samples. *Blocks* are groups of 100 numbers with the same first two digits. For example, 555-444-12XX would be considered one 100-block listed on the sampling frame. Once a household is found in an area code + exchange + block, all of the numbers in the particular block are called, dramatically reducing the number of nonhousehold calls, as is done in the Mitofsky-Waksberg approach to RDD sampling.

Importance of Well-Designed Probability Surveys

The elements of design, including stratification, clustering, and statistical weights, should not be ignored in analyzing the results of a probability sample survey. Ignoring the sampling design may underestimate variability, affecting potential inferences. Software has advanced greatly in recent years and has become more accessible to researchers wishing to estimate population characteristics and their associated standard errors using sample survey data. This should encourage researchers to carry out well-designed probability surveys to attain their study objectives and to use appropriate methods to analyze their data.

Erinn M. Hade and Stanley Lemeshow

See also Census; Cluster Sample; Coverage Error; Elements; Finite Population; Mitosky-Waksberg Sampling; Multi-Stage Sample; *n*; *N*; Nonprobability Sample; Probability Proportional to Size (PPS) Sampling; Purposive Sample; Quota Sampling; Random; Random-Digit Dialing (RDD); Random Sampling; Replacement; Sampling Frame; Simple Random Sample; Snowball Sampling; Strata; Stratified Sampling; Systematic Sampling; Target Population; Unbiased Statistic; Unit of Observation; Validity; Variance

Further Readings

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications* (3rd ed.). New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury.

Probable Electorate

The *probable electorate* is defined as those citizens who are registered to vote and who very likely will vote in an upcoming election. In election polling and surveying, this concept is operationalized as a sample of pre-election survey respondents whose candidate preferences have been weighted by the respondents' estimated likelihood of voting.

One ongoing challenge in election polling lies in ascertaining which respondents will actually turn out to vote. The distribution of candidate support among the sample in the aggregate, while revealing, is essentially irrelevant when it comes to making predictions about election outcomes. What matters are the choices of those respondents who will actually vote on (or in some cases before) Election Day. If survey researchers had a crystal ball, they could determine with precision which respondents would actually be voters (but then, of course, they could dispense with the pre-election survey altogether and simply foretell the outcome of the election). In the real world, techniques are necessary to carve out a probable electorate upon which to base pre-election predictions.

Estimating a probable electorate requires either the creation of a binary turnout screener, by which survey respondents are assigned voting probabilities of either 0 or 1, or the use of a weighting method in which respondents are assigned estimated probabilities on a continuum that can range anywhere between 0 and 1. However, few pollsters provide detailed information on how, exactly, they construct their likely voter estimates, treating their procedures like a secret (proprietary) formula. What is known about how probable electorates are constructed in practice is based on a relatively slim store of publicly available information.

Turnout screening (i.e., a series of survey questions that helps the researcher estimate the likelihood a given respondent will vote) aims to differentiate voters and nonvoters, taking into account the candidate preferences of the former, but not the latter, when making election predictions. Turnout screens can take multiple forms, some simple, some exceedingly complex. On the simple side, a screen might be based on a self-reported turnout intention, asking respondents how likely they are to vote in an upcoming election, then counting only the preferences of those who say they will "definitely" or "probably" vote.

Of course, not everyone who says they plan to vote will actually do so (although most of those who say they will not, do not). More sophisticated screeners take into consideration registration status, past voting behavior, campaign interest, knowledge about the location of polling places, or some combination. Some turnout screeners are based on as many as a dozen questions or more, although research suggests that a screener that includes 15 questions is not much more accurate than screeners based on eight or fewer items.

Using multi-item indices to construct a turnout screening method produces estimated levels of voting likelihood (a "turnout score") that require the application of some threshold or cut-point: for example, should respondents scoring a 6 or above on a 10-point likelihood scale be counted as probable voters, or only those scoring 7 or higher? The answer depends in part on a priori estimates of what turnout will be on Election Day. Higher anticipated turnout entails including more respondents in one's probable electorate.

An alternative to imposing a threshold in order to create a binary screener involves assigning turnout probabilities (ranging from 0 to 1, like a propensity score) for every respondent in one's sample. Doing so requires deriving estimates from out-of-sample models, usually involving "validated" reports of voter turnout. Such models include demographic, behavioral, and in some cases attitudinal factors and yield weights that can be applied to respondents in any given pre-election survey that includes the same questions.

Once a respondent's voting probability has been calculated, his or her candidate preference is simply weighted accordingly. The preferences of those deemed more likely to vote are thus counted more than those of respondents with little chance of turning out. Respondents who say they do not intend to vote, or those who are not registered, can be assigned a probability of 0. For others, though, this method recognizes that even those unlikely to vote have some chance of casting a ballot while acknowledging that not all "likely voters" are created equal.

Paul Freedman

See also Election Polls; Likely Voter; Pre-Election Polls; Propensity Scores

Further Readings

Crespi, I. (1988). *Pre-election polling: Sources of accuracy and error*. New York: Russell Sage Foundation.

Daves, R. (2000). Who will vote? Ascertaining likelihood to vote and modeling a probable electorate in preelection polls. In P. Lavrakas & M. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 206–223). New York: Chatham House.

Perry, P. (1960). Election survey procedures of the Gallup Poll. *Public Opinion Quarterly 24*, 531–542.

Probing

Probing involves the use of specific words or other interviewing techniques by an interviewer to clarify or to seek elaboration of a person's response to a survey question. When respondents provide incomplete or irrelevant answers to survey questions, it becomes necessary for interviewers to query respondents further in an effort to obtain a more complete or specific answer to a given question. Although survey researchers may take great care in constructing questionnaires in terms of the wording of both questions and response options, some respondents may not provide a response in the format pre-specified by the researcher, especially when answering closed-ended questions. Likewise, respondents may offer vague or overly simplistic replies to open-ended questions, that is, questions with no predetermined response categories. Additionally, respondents may simply provide the response, "I don't know," which the researcher may or may not want to accept as a valid response.

The following are examples of responses requiring probing:

> Interviewer question: *In the past 12 months, has a doctor, nurse, or other health professional given you advice about your weight?* [The valid response options are "Yes" or "No."]
>
> *Irrelevant respondent answer:* "My husband is on a diet."
>
> *Unclear respondent answer:* "People are always telling me I need to gain some weight."

To elicit an acceptable response to a given item, interviewers should use an appropriate neutral, unbiased, or "nondirective" probe. Interviewers should take care to elicit the information needed without suggesting or leading the respondent to a particular response. Researchers should include suggested wording for potential probes in their questionnaires, in an effort to equip interviewers to handle these situations. Although every situation is different, there are some basic techniques and general rules that most interviewers find successful when probing respondents for information.

1. *Repeat the question.* When respondents appear to have misunderstood or misinterpreted a survey question, repeating the question is the best probe. This technique is used with the expectation that after hearing the survey question a second time, the respondent will understand what information the question is intended to collect.

2. *Silent probe.* Interviewers may also use a silent probe, which is a pause or hesitation intended to indicate to a respondent that the interviewer may be waiting for additional information or clarification on a response. Oftentimes, interviewers will utilize this technique during later stages of an interview, once a respondent's habits or response patterns have become more obvious.

3. *Neutral question or statement.* When a respondent offers an inadequate response, interviewers use neutral questions or statements to encourage a respondent to elaborate on their initial response. Examples of good neutral probes are "What do you mean?" "How do you mean?" "Please tell me what you have in mind," "Please tell me more about . . . " Note that these probes maintain neutrality and do not lead the respondent by asking things such as "Do you mean you don't support the bill?"

4. *Clarification probe.* When a response to an item is unclear, ambiguous, or contradictory, interviewers will use clarification probes. Examples of good clarification probes are "Can you give me an example?" or "Could you be more specific?" Whereas these probes are helpful, interviewers must be careful not to appear to challenge the respondent when clarifying a statement. Interviewers must know when to draw the line between probing a respondent for more or better information and making the respondent feel pressured to answer an item, which could lead to an outright refusal to continue participation in the rest of the survey.

Most interviewers will agree that "I don't know" is the response to survey items requiring probing that occurs most frequently. Because a "don't know" response is vague and can mean any number of things, interviewers are often challenged with the need to choose the correct probe. Interviewers are also challenged to not cross the line between probing and pressuring in this situation as well. Interviewers are trained to remind respondents that participating in a survey is not a test, and that there are no right and wrong answers. A good practice for interviewers is encouraging a given respondent who provides a "don't know" response to an item to give his or her "best estimate," or "best guess." Encouraging the respondent to answer an item and reminding them that their answers, no matter what they are, are the "right answers" to a given question will likely lead to better consideration of the question and a higher quality response.

David James Roe

See also Bias; Closed-Ended Question; Interviewer Effects; Interviewer Neutrality; Interviewer-Related Error; Nondirective Probing; Open-Ended Questions

Further Readings

Weisberg, H. F., Krosnick, J., & Bowen, B. (1996). *An introduction to survey research, polling and data analysis* (3rd ed.). Thousand Oaks, CA: Sage.

Propensity Scores

Propensity scoring was developed as a statistical technique for adjusting for selection bias in causal estimates of treatment effects in observational studies. Unlike randomized experiments, in observational studies researchers have no control over treatment assignment, and, as a result, individuals who receive different treatments may be very different in terms of their observed covariates. These differences, if left unadjusted, can lead to biased estimates of treatment effects. For example, if smokers tend to be older than nonsmokers, then comparisons of smokers and nonsmokers will be confounded with age. Propensity scores can be used to adjust for this observed selection bias.

Survey researchers have used propensity scores to adjust for nonresponse bias, which arises when respondents and nonrespondents differ systematically in terms of observed covariates, and to adjust for selection (coverage) bias, which arises when some of the population is systematically omitted from the sample.

Estimating Treatment Effects in Observational Studies

In the context of estimating causal effects in observational studies, the *propensity score* is the conditional probability that an individual belongs to a specific treatment group (e.g., the treated group or the control group) given a set of observed covariates. Propensity scores are balancing scores, meaning that within subclasses that are homogeneous in the propensity score, the distributions of the covariates are the same for treated and control units (i.e., are "balanced"). This makes it possible to estimate treatment effects, controlling for covariates, by matching or subclassifying on the propensity score since comparisons of individuals with different treatments made within these matched pairs or groups are not confounded by differences in observed covariate distributions. An unbiased estimate of the average treatment effect is obtained when researchers have controlled for all relevant covariates (the so-called strongly ignorable treatment assignment assumption).

Propensity scores are usually estimated by logistic regression, although other models can be used. Most of the propensity score literature focuses on the binary treatment case (e.g., treated vs. control); however, propensity score methods have been extended more recently to accommodate multiple treatment groups. Propensity scores can be used for stratification, matching, or as a covariate in future analyses.

Adjusting for Nonresponse Bias in Surveys

Propensity scores have been used in survey research to adjust for nonresponse bias. In this case, the propensity score is the probability of being a respondent given observed characteristics. These propensity scores can be used in post-stratification, weighting adjustments, and imputation.

Post-stratification using propensity scores is useful when there are a substantial number of variables available for post-stratification, as there might be in panel surveys where information from a previous wave is available for nonrespondents. In this situation, standard post-stratification methods that form weighting cells by complete cross-classification of all the control variables are not practical, since the number of weighting cells would be very large and could include cells with nonrespondents but few or no respondents. If there are no respondents in a cell, the nonresponse weight adjustment is infinite. An alternative is to estimate the propensity to be a respondent using the available observed variables and then to group these estimated propensities into a reasonable number of weighting classes. This takes advantage of the propensity score's ability to adjust for a large number of covariates simultaneously using only a single scalar summary (the propensity score).

Propensity scores can also be used directly in nonresponse weighting adjustments. To adjust for nonresponse, survey weights can be multiplied by the inverse of the estimated propensity score for that respondent. This avoids the assumptions involved with grouping together respondents with similar but not identical response propensity scores into weighting classes. However, this method relies more heavily on the correct specification of the propensity score model, since the estimated propensities are used directly in the nonresponse adjustment and not just to form weighting cells; and this method can produce estimators with very large variances, since respondents with very small estimated propensity scores receive very large weights.

Propensity scores can also be used for imputation for nonresponse through propensity-matching techniques. The idea here is similar to that of *nearest neighbor imputation* or *predictive-mean-matching imputation*. Nonrespondents are matched to respondents with similar propensity scores, and then these respondents serve as "donors" for the nonrespondents' missing information. This can be done using single or multiple imputation. For single imputation, the respondent who is closest in terms of estimated propensity score can be chosen as the donor or can be selected at random from a pool of donors who are all relatively close. For multiple imputation, a pool of potential donors can be created by specifying tolerances for the difference in propensity scores between the nonrespondent and the respondents, or by grouping together the nearest k respondents, or by stratifying on the estimated propensity scores. The approximate Bayesian bootstrap can then be used to create "proper" multiple imputations that correctly approximate the uncertainty in the imputed values. To do this, a bootstrap sample of potential donors is selected at random with replacement from the available pool of potential donors for each nonrespondent. Imputation donors for each nonrespondent are then selected at random with replacement from this bootstrap sample. This process is repeated to create multiple imputations for each nonrespondent.

Adjusting for Selection Bias in Surveys

Propensity scores have also been used in survey research to adjust for selection (coverage) bias. For example, propensity scores have been used to adjust for the selection bias that arises in Web surveys because not all members of the target population have Internet access. For this approach to be successful, it is necessary to have a *reference survey* that does not have the selection bias problems of the Web survey. For example, the reference survey could be a telephone survey, assuming that the entire population of interest has telephone service. The reference survey and the Web survey are combined to estimate the propensity to be a Web respondent (as opposed to a telephone respondent), given observed covariates. The reference survey can be a very short survey, containing only questions about characteristics, attitudes, or behaviors that are hypothesized to differ between the Web population and the telephone population. The resulting estimated propensities can be used for post-stratification as described previously, except that instead of each cell containing respondents and nonrespondents, each cell contains Web respondents and telephone respondents. Weight adjustments are made so that the Web survey's weighted proportion of respondents in each cell matches the reference survey's estimated proportion in each cell.

Elaine L. Zanutto

See also Bootstrapping; Coverage Error; Experimental Design; Imputation; Multiple Imputation; Nonresponse Bias; Panel Survey; Post-Stratification; Propensity-Weighted Web Survey; Wave; Weighting

Further Readings

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–2281.

Lavori, P. W., Dawson, R., & Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, *14*, 1913–1925.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, *54*, 139–157.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387–394.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Schonlau, M., Zapert, K., Payne Simon, L., Haynes Sanstad, K., Marcus, S., Adams, J., et al. (2004). A comparison between responses from a propensity-weighted Web survey and an identical RDD survey. *Social Science Computer Review*, *22*, 128–138.

Propensity-Weighted Web Survey

Because Web surveys are often convenience samples, traditional methods for statistical inference do not readily apply. Propensity scoring is one attempt to correct for the selection bias of a nonrandom Web sample. Broadly speaking, the propensity scoring adjustment is accomplished by reweighting the convenience sample such that the distribution of so-called propensity variables corresponds to that of a reference sample. *Propensity variables* (or *propensity questions*) can be any survey questions that have been answered both by the Web survey participants and by the respondents of the reference sample. *The reference sample* is a separate probability sample (e.g., a random-digit dialed [RDD] phone survey) from a possibly much shorter survey that contains only the propensity questions.

History

Propensity scoring has traditionally been applied in biostatistics to estimate causal effects. Harris Interactive, a New York–based Web survey business, first applied propensity scoring to correct for selection bias in Web surveys. Harris Interactive uses special "Webographic" questions as propensity variables. Webographic questions—also called "lifestyle," "attitudinal," or "psychographic" questions—are meant to capture the difference between the online and the offline population.

The Practice of Propensity Scoring

In practice, the Web sample and the reference sample are combined to form a single data set. An indicator variable, indicating whether a respondent is from the Web sample, is regressed on the propensity questions, usually via logistic regression, representing the probability that respondents within certain characteristics are from the Web sample. The propensity scores are computed as the predicted values from this logistic regression. The propensity scores can be used in a variety of ways, including as weights for stratification and matching techniques (research). Using the inverse propensity scores as adjustment weights is appealing because the concept of reweighting is familiar and because standard survey software can be used to conduct statistical tests. Another popular method is to stratify the propensity scores into quintiles and to use the resulting five-level categorical variable in a post-stratification step.

Limitations

First, it is not possible to adjust for unbalanced, unobserved variables that correlate with outcomes unless the unobserved variable strongly correlates with observed variables. Second, to calibrate the Web survey, the propensity scoring approach requires a reference sample. For example, Harris Interactive conducts regular RDD phone surveys for that purpose. This requirement currently limits the appeal of this method and makes it most useful for panel surveys. Propensity scoring attempts to achieve balance. That is, after the propensity weighting adjustment, the distribution of the propensity variables should be the same for both the Web sample and the reference sample. The traditional goal to find a logistic regression model that fits the data well is therefore not necessarily useful. Instead, the researcher should verify whether balance was achieved. Preliminary research seems to indicate that the propensity adjustment reduces the bias considerably but does not remove it altogether for all outcome variables. One direction for future research is to find out which set of propensity variables works best for which set of outcomes. Of additional note, propensity scoring has also been applied to adjust for nonresponse bias when data are available for both nonrespondents and respondents of a survey.

Matthias Schonlau

See also Convenience Sampling; Nonresponse Bias; Post-Stratification; Probability Sample; Propensity Scores; Psychographic Measures

Further Readings

Schonlau, M., Van Soest, A., Kapteyn, A., Couper, M., & Winter, J. (in press). Selection bias in Web surveys and the use of propensity scores. *Sociological Methods and Research.*

Schonlau, M., Zapert, K., Payne Simon, L., Sanstad, K., Marcus, S., Adams, J., et al. (2004). A comparison between a propensity weighted Web survey and an identical RDD survey. *Social Science Computer Review*, *22*(1), 128–138.

Proportional Allocation to Strata

Proportional allocation is a procedure for dividing a sample among the strata in a stratified sample survey. A *sample survey* collects data from a population in order to estimate population characteristics. A *stratified sample* selects separate samples from subgroups of the population, which are called “strata” and can often increase the accuracy of survey results. In order to implement stratified sampling, it is necessary to be able to divide the population at least implicitly into strata before sampling. Given a budget that allows gathering data on n subjects or a budget amount $\$B$, there is a need to decide how to allocate the resources for data gathering to the strata. Three factors typically affect the distribution of resources to the strata: (1) the population size, (2) the variability of values, and (3) the data collection per unit cost in the strata. One also can have special interest in characteristics of some particular strata that could affect allocations.

Assuming the goal of the survey is to estimate a total or average for the entire population, the variability of values are not known to differ substantially by strata, and data collection costs are believed to be roughly equal by strata, one could consider allocating sample size to strata proportional to strata population sizes. That is, if there are H strata with population size N_h in stratum h, $h=1,2,\ldots,H$, and one can afford to collect data on n units, then the proportional allocation sample size for stratum h is $n_h=n(N_h/N)$, where $N=\sum_{h=1}^{H} N_h$ is the total population size. As a result, strata with large numbers of units in their populations receive more sample, whereas small strata receive less sample. With roughly equal per-unit data collection costs, a budget amount $\$B$ corresponds to a total sample size n. If the n_h's are not integer, then one must round the numbers to integers for sample selection. Rounding does not necessarily move all values to the closest integer for all strata, because the total sample size n needs to be allocated.

Suppose you want to collect data on students at a large public university. Questions of interest could be hours worked per week, amount expended per semester on textbooks, amount of time spent eating at restaurants in a week, number of trips to the airport in a semester, and whether or not friends smoke cigarettes. The students selected for the survey could be contacted via their university email addresses and asked to complete an online Web survey. A survey can be preferable to contacting every student, because better efforts can often be made for a sample to encourage response and check data quality. Administrative records contain college year designations (1st, 2nd, 3rd, 4th) for each student in the target population; college years can be used as strata. Suppose the total sample size is allowed to be 1,600 students. Equal allocation to strata would sample 400 students from each year. Table 1 presents proportional allocations of students to the four strata based on total enrollments by college year; these numbers are similar to 2006 enrollment at Iowa State University. As can be seen in the table, the stratum of fourth-year students receives the largest sample ($n_4=503$), where as the stratum of second-year students receives the smallest ($n_2=324$).

Proportional allocation of the sample in stratified sampling can yield more precise estimates of means (e.g., average amount spent on textbooks) and totals (e.g., number of trips to the airport) of characteristics in the population than *equal allocation* to strata with the sample total sample size (n) when strata are of very unequal sizes. In fact, if data collection costs per unit and the variance of values are the same across strata, then proportional allocation is optimal under *stratified simple random sampling*. Suppose 1,600 students are to be chosen from the population for a survey, and the college (Agriculture, Business, Design, Engineering, Human Sciences, and Liberal Arts & Sciences) of enrollment within the university is used to stratify students into groups. Table 2 presents an example for which some strata would have much larger or smaller samples than others.

Table 1 Proportional allocation of 1,600 students to four strata defined by college year

Year	*Population Size: Total Enrollment*	*Sample Size: $n_h=nN_h/N$*	*Sample Size: Rounded Values*
First	5,000	399.4	399
Second	4,058	324.1	324
Third	4,677	373.6	374
Fourth	6,296	502.9	503
Total	20,031	1,600.0	1,600

Table 2 Proportional allocation of 1,600 students to six strata defined by college

College	*Population Size: Total Enrollment*	*Sample Size:* $n_h = nN_h/N$	*Sample Size: Rounded Values*	*Sample Size: Adjusted to 200 Minimum*
Agriculture	2,488	198.7	199	200
Business	3,247	259.4	259	246
Design	1,738	138.8	139	200
Engineering	4,356	347.9	348	331
Human Sciences	2,641	211.0	211	201
Liberal Arts & Sciences	5,561	444.2	444	422
Total	20,031	1,600.0	1,600	1,600

If variances in strata are quite different or costs of data collection vary by strata, then one could consider *optimal allocation* and probably produce more precise estimates of population means and totals. If one plans to make inferential statements about all strata separately, then it might be a good idea to adjust sample sizes to be closer to equal than suggested by proportional allocation. The right-most column of Table 2 presents a compromise between proportional and equal allocation to strata. The minimum sample size per stratum was increased to 200, and the remaining sample sizes were ratio adjusted so that the full sample size is still 1,600. In this example, one could, of course, stratify by college year and college simultaneously, thereby producing 24 strata.

Michael D. Larsen

See also Disproportionate Allocation to Strata; Optimal Allocation; Stratified Sampling

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Protection of Human Subjects

In most instances, survey research requires human participation and the self-report of opinions, facts, behaviors, and/or experiences. Some survey research also requires the collection of physiological samples. It may also involve the use of administrative data if the researcher has access to Social Security records, health records, military service information, and the like. Past abuses of human research participants have led to the development of a set of ethical guidelines to which researchers must adhere in order to protect human subjects from physical, mental, and emotional harm.

Historical Perspective on the Abuse of Human Subjects in Research

One of the earliest examples of the abuse of human subjects was the infamous Tuskegee Syphilis Study that was conducted from 1932 to 1972 in Macon County, Alabama. The purpose of the study was to observe the progression of untreated syphilis in black men. A total of 399 black men with syphilis and 201 black men without it were recruited. The men were told that they were receiving treatment when in fact they were not. In addition, the physicians in the county were told not to treat these men since it would interfere with the study. When penicillin became the established treatment for syphilis in 1947, the treatment was withheld from the men in the study known to have syphilis. The men and the course of their illnesses were documented and observed until they died. Their spouses caught the disease from them and children were born infected with it. Yet their diagnosis and treatment was withheld. Finally, a front-page article in *The New York Times* exposed what was going on, and the public became outraged. Consequently, a panel was appointed by the Department of Health and Scientific Affairs to review the study. Although the men had freely consented to participate, the panel discovered that they were never told of the considerable risks to their health and to the health of their families. In October 1972, the panel advised that the study be discontinued because the men had not been told about the risk of physical harm, which was necessary for the men to have given informed consent. One month later, the Assistant Secretary of Health and Scientific Affairs announced that the study was ended. The termination of the study was followed by a lawsuit brought by the National Association for the Advancement of Colored People (NAACP), which resulted in an award of $9 million to the participants as well as free medical care and burials for those who were still living.

A second example of the abuse of human research participants was the Willowbrook Hepatitis Study that was carried out from 1963 through 1966 at the Willowbrook State School of New York. Here, mentally disabled individuals were intentionally infected with hepatitis in order to study the disease and a possible treatment, gamma globulin. These individuals were all children who by reason of their mental status could not legally give consent themselves. Parents who wished to have their children admitted to the institution were told that there was no room for their child unless they agreed to have the child participate in the hepatitis studies. The investigators claimed that most of the children at Willowbrook would eventually become infected with hepatitis anyway. The public became outraged that the parents and their children were forced to enter the hepatitis research program or else admission to the institution would be denied.

There were also numerous studies that involved radiation in order to discover its effects; however, in some of these studies participants were not warned about the potential risk of cancer. For example, some of the studies involved prison inmates at Oregon State Prison and Washington State Prison who received radiation to the testicles but were not warned of the risk of testicular cancer; these studies took place from 1963 to 1971. In the late 1940s, the impact of radiation on fetal development was studied at the Vanderbilt University; 800 pregnant women were exposed to radiation. In 1998, the participants were awarded a large settlement.

In addition, there were the atrocities committed by Nazis such as Dr. Josef Mengele at Auschwitz, who was infamous for his "research" on twins. For example, the Nazis prized blonde hair and blue eyes; therefore, a major focus of Mengele's work was to experiment with changing hair and eye color. On Mengele's orders, chemicals were put in children's eyes with eye drops or needles in an attempt to change their eye color. Afterward, the children could not see for several days. Since the children were incarcerated in a death camp, they had no choice but to submit to the painful experiments.

These cases highlight a number of the issues that needed to be addressed in the creation of a code to protect human research participants.

Codes of Research Ethics

The first attempt at a code of research ethics was the Nuremberg Code, and it was created in response to the war crimes committed by the Nazis. The 10 points of the Nuremberg Code are paraphrased in Table 1.

Table 1 Summary of the Nuremberg Code

1. Participation must be completely voluntary and the participant must have the legal capacity to consent. The person must be fully informed of the nature and duration of the research.
2. The research must be of benefit to society and the findings cannot be obtained by any other method.
3. The research should have a sound foundation in animal research and with knowledge of the history of the problem under study.
4. Unnecessary physical or psychological harm must be avoided.
5. No research shall be performed that may cause death or disability to the participant.
6. The degree of risk must not exceed the expected benefit of the research.
7. Proper plans must be made and adequate facilities must be provided to protect the research participant from even remote possibilities of injury, death, or disability.
8. Research may only be conducted by highly qualified, trained scientists.
9. During the research, the participant is free to stop participating at any time.
10. The researcher must be prepared to end the experiment if it becomes clear to the researcher that continuing will harm the participants.

This laid the foundation for most of the current ethical standards in research. The American Psychological Association (APA) began writing ethical guidelines in 1947, and its first ethical code was published in 1953. Numerous revisions of the APA ethical guidelines for the protection of human participants have taken place over the years, and the latest revision occurred in 2002. Note that APA guidelines address more issues than the Nuremberg Code did; one difference is the issue of research with children who are not yet old enough to give legal consent to participate. Table 2 provides a summary of the current APA ethical guidelines for the protection of human participants.

Carla R. Scanlan

See also Certificate of Confidentiality; Debriefing; Deception; Ethical Principles; Informed Consent; Voluntary Participation

Table 2 Summary of the 2002 American Psychological Association guidelines for the protection of human research participants

1. Research proposals submitted to institutional review boards (IRB) must contain accurate information. Upon approval, researchers cannot make changes without seeking approval from the IRB.
2. Informed consent is usually required and includes: (a) the purpose of the research, expected duration, and procedures; (b) the right to decline to participate and to withdraw once the study has begun; (c) the potential consequences of declining or withdrawing; (d) any potential risks, discomfort, or adverse effects; (e) any potential research benefits; (f) the limits of confidentiality; (g) incentives for participation; (h) whom to contact for questions about the research and participants' rights. Researchers provide opportunity for the participants to ask questions and receive answers.
3. When research is conducted that includes experimental treatments, participants shall be so informed at the beginning of the study of (a) the experimental nature of the treatment, (b) the services that will or will not be available to the control group(s) if appropriate; (c) the means by which assignment to treatment or control group is made; (d) available treatment alternative if the person does not want to participate or withdraws after the study has begun; (e) compensation for costs of participating.
4. Informed consent shall be obtained when voices or images are recorded as data unless (a) the research consists only of naturalistic observations taking place in public places and the recording is not anticipated to cause harm; or (b) the research design includes deception, and consent for the use of the recording is obtained during debriefing.
5. When psychologists conduct research with clients/patients, students, or subordinates as participants, steps must be taken to protect the participants from adverse consequences of declining or withdrawing from participation. When research participation is a course requirement, or an opportunity to earn extra credit, an alternative choice of activity is available.
6. Informed consent may not be required where (a) research would not be reasonably expected to cause distress or harm and (b) involves (i) the study of normal educational practices, curricula, or classroom management methods in educational settings; (ii) only anonymous questionnaires, archival research, or naturalistic observations that would not place participants at risk of civil or criminal liability nor damage financial standing, employability, or reputation, and confidentiality is protected; or (iii) the study of organizational factors conducted in the workplace poses no risk to participants' employability and confidentiality is preserved.
7. Psychologists make reasonable efforts to avoid offering excessive or inappropriate financial or other inducement for research participation that would result in coercion.
8. Deception in research shall be used only if it is justified by the study's significant prospective scientific value and nondeceptive alternatives are not feasible. Deception shall not be used if it will cause physical pain or severe emotional distress.
9. (a) Participants shall be offered promptly supplied information about the outcome of the study; (b) if delay or withholding of the study outcome is necessary, reasonable measures must be taken to reduce the risk of harm; (c) when researchers realize that research procedures have harmed a participant, they take reasonable steps to minimize the harm.

Further Readings

The Belmont Report: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm
Bordens, K. S., & Abbott, B. B. (2005). *Research design and methods* (6th ed.). New York: McGraw-Hill.
Cases of Unethical Research: http://www.clarion.edu/academic/adeptt/bpcluster/cases.htm
Institutional Review Board Guidebook: http://www.hhs.gov/ohrp/irb/irb_appendices.htm
Lagnado, L. M., & Dekel, S. C. (1991). *Children of the flames*. New York: Morrow.
The Tuskegee Timeline: http://www.cdc.gov/nchstp/od/tuskegee/time.htm
Willowbrook Hepatitis Study: http://hstraining.orda.ucsb.edu/willowbrook.htm

Proxy Respondent

If a respondent reports on the properties or activities of another person or group of persons (e.g., an entire household or a company), the respondent is said to be a *proxy* respondent. In some cases, proxy responses may be a part of the design of the survey. In the U.S. Current Population Survey, for example, a single responsible adult is asked to report for all members of the household 14 years of age or older. In many surveys, adults are asked to report for children. In other cases, proxy responses are used only when there is a particular reason that the targeted person cannot report. In household travel surveys, for example, data are typically sought directly from each adult member of the household. Only if a particular adult is on extended travel, has language or medical difficulties that would interfere with responding, or some similar reason, would a proxy respondent be used.

Since a proxy response is treated the same as a self-reported response, an obvious benefit of allowing proxy responses is to increase the response rate. If the targeted person is unavailable for the entire survey period, the only way to get a response may be to accept a proxy response. In surveys in which information is sought about all members of a household, the lack of a response from one member, if only self-reported data are permitted, could jeopardize the utility of the information from the others.

Not allowing proxy responding also may increase nonresponse bias. Those unavailable to respond for themselves are more likely to be on long trips, in the hospital, or away at college, and so on, than those who are available. If these factors are related to the purpose of the survey, then the use of proxy respondents should be considered.

Because the proxy respondent will have a different perspective and set of memories than the targeted person, proxy responding can be expected to affect measurement error in the survey. Proxy respondents may record fewer less salient events (e.g., smaller purchases or short trips) than respondents reporting for themselves. On the other hand, there are instances when proxy responses may be more accurate than reports from self-respondents, for example, the main record keeper in a household reporting for other household members.

Some survey items do not lend themselves to proxy responding because measurement error is apt to be particularly great. A noteworthy case is attitudinal items. Even if the proxy respondent knows the targeted person extremely well, the attitudes of the proxy respondent will likely be confounded with the attitudes of the targeted person in the responses.

Michael P. Cohen

See also Eligibility; Measurement Error; Nonresponse; Nonresponse Bias; Respondent-Related Error; Response Rates; Unit; Unit Nonresponse

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Pseudo-Polls

The term *pseudo-poll* refers to a number of practices that may appear to be legitimate polls but are not. A legitimate poll uses scientific sampling to learn about the opinions and behaviors of a population. Pseudo-polls include unscientific (and thus, unreliable) attempts to measure opinions and behaviors as well as other practices that look like polls but are designed for purposes other than legitimate research.

A variety of techniques are used to conduct unscientific assessments of opinion, all of which are considered pseudo-polls. These can be used by media and other organizations as an inexpensive way to measure public opinion and get the audience involved. However, they are very problematic from a data quality standpoint and should not be referred to as polls.

One common approach is the *call-in poll*. This is a method used when audience members or newspaper readers are invited to call a phone number and register their opinions. This is also to referred as a *SLOP*, a term coined by survey researcher Norman Bradburn, which stands for "self-selected listener opinion poll." A similar technique is used by newspapers and magazines that publish questionnaires and ask their readers to fill them out and mail them in. A newer approach, called a *log-in poll*, is where Web sites post questions for people to click on to register their opinions, with the aggregate results being displayed instantly.

A major problem with all of these types of pseudo-polls is that the participants are entirely self-selected. Only those people who tuned in to that particular broadcast at that time or read that newspaper or went to that Web site can be included. Further, those who make the effort to participate are often very different from those in the audience who do not. Those who participate are often more interested in the topic or feel very strongly about it.

Another big problem with pseudo-polls is that they are open to manipulation by those with a vested interest in the topic. With call-in polls there is no limit on the number of calls that can be placed, so people can (and do) call multiple times and groups can set up elaborate operations to flood the phone line in support of their point of view. Manipulation is also a problem with Internet log-in polls. For example, multiple clicks can be possible, and it is common for political campaigns to direct their supporters to Web sites that have log-in polls to boost their position's and/or candidate's standing.

For these reasons, these types of pseudo-polls often produce biased results and should be ignored. Legitimate pollsters who are concerned with poll accuracy avoid these types of bias by selecting respondents using probability sampling techniques. Although survey researchers know that these pseudo-polls should not be taken seriously, many members of the public do not make a distinction between these and the real thing. In fact, pseudo-polls may be incorrectly seen as even more credible than real polls because they often have much larger sample sizes.

There are other types of pseudo-polls for which the purpose is not opinion collection. One of these is called a "push poll." This is not a poll at all, but rather a sneaky and unethical telemarketing trick used by some candidates' political campaigns. Large numbers of people are called and are asked to participate in what appears to them to be a legitimate poll. In the process people are told negative, often false, things about the opposing candidates. The purpose is to shift or "push" people's opinions about the candidates rather than to collect data.

Sales and fundraising calls designed to look like polls are other types of unethical pseudo-poll not designed to collect data. SUGing, which stands for "selling under the guise" of survey research, takes people through what appears to be a legitimate poll to try to get them to purchase something. The information obtained in the purported survey later may be used to hone the ultimate sales pitch.

FRUGing is similar and stands for "fundraising under the guise" of survey research. Respondents are administered a bogus survey whose purpose is to obtain donations for the given cause. Often the questionnaire is targeted to those with a certain point of view, and the question wording is biased in favor of that point of view in an effort to increase the respondent's chances of contributing.

The purpose of legitimate polling and survey research is to provide accurate information about a population's opinions and behaviors. Call-in polls and similar efforts using mail, Internet, or fax may attempt to provide this information but fail because the procedures they employ yield biased data. Other types of pseudo-polls, such as push polls and SUGing and FRUGing efforts, are deceptive practices designed to manipulate people and are condemned by professional survey research associations.

Daniel M. Merkle

See also Call-In Polls; 800 Poll; FRUGing; Log-In Polls; 900 Poll; Probability Sample; Push Poll; Self-Selected Listener Opinion Poll (SLOP); Self-Selected Sample; Self-Selection Bias; SUGing; Survey Ethics; Telemarketing

Further Readings

Gawiser, S. R., & Witt, G. E. (2002). *20 questions journalists should ask about poll results.* Retrieved June 12, 2008, from http://www.ncpp.org/?q=node/4

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Psychographic Measure

A psychographic measure is a variable that represents a personal characteristic of an individual that is not a physical trait, as are demographic measures (age, gender, height, etc.). Rather, psychographic variables include personality traits, lifestyle preferences or interests, values or beliefs, and attitudes or opinions.

Because psychographics are not directly observable, as are many demographics, nor do they have a "factual" basis as do demographics, their measurement is less precise. Surveys are routinely used to measure psychographics, and they can serve as powerful independent variables in helping explain many dependent variables of interest. For example, in political science, understanding why people vote for the presidential candidates they do is heavily explained by psychographics such as party affiliation, socioeconomic class, religion, veteran status, sexual orientation, and other values, beliefs, and attitudes.

Psychographic characteristics often are measured by scales formed from a set of Likert items in which a series of statements are made and the respondent indicates the extent to which she or her agrees or disagrees with each statement. Factor analysis and other scaling techniques can then be used to identify the most reliable set of items to make up a scale for a particular psychographic measure.

Including psychographic measures in a questionnaire often increases the richness of one's survey database because they often are powerful predictors of other variables of interest. Psychographics are used routinely by market researchers. There are myriad scales that measure various psychographics; many are in the public domain, while others are highly proprietary and are quite expensive to access. In the latter case, many companies have devised and validated proprietary psychographic scales that require a researcher to send her or his raw survey data to the company. Then, for a fee, the company "scores" the raw data, thereby creating the psychographic variables for each respondent the researcher has in her or his data set.

Paul J. Lavrakas

See also Attitudes; Demographic Measure; Likert Scale; Opinions

Further Readings

Goldberg, M. (1976). Identifying relevant psychographic segments: How specifying product functions can help. *The Journal of Consumer Research*, *3*(3), 163–169.

Kahle, L., & Chiagouris, L. (1997). *Values, lifestyles, and psychographics.* Mahwah, NJ: Lawrence Erlbaum.

Public Opinion

Public opinion is one of the oldest and most widely invoked concepts in the social sciences and has many meanings and controversies associated with it. The concept of public opinion has evolved during the past three millennia as the notion of democracy evolved. Public opinion plays a central role in democracy, and, as such, the ability to "measure" public opinion accurately has become a critical endeavor for modern-day democracies.

Historical and Philosophical Background

The concept has origins in early political philosophy dating to the writings of Plato, who was mainly concerned about the mercurial nature of crowds and the problems associated with direct democracy. Aristotle is often cited on the point of the collective wisdom of the group being wiser than individuals', but he was not in general a reliable advocate for rationality of the common people.

The controversies about the competence and rationality of people continued throughout the history of Western philosophy, through the Middle Ages to the present day. From the perspective of our contemporary representative democracies, public opinion is considered with reverence as "the will of the people." But such representations are often accompanied by doubts about the extent of the public's competencies, levels of knowledge, ability to weigh contradictory information, and level of emotion.

The origin of public opinion is often portrayed as an achievement of Western culture and civilization, coming together in word and deed on the eve of the French Revolution. Informed people met frequently in the upper-class salons of the day to discuss public affairs. Records of this era are clear about the frequency and quality of these conversations and their importance in influencing events in that time and place. Coming together were the developments in popular information, regular venues for conversation and deliberation of the issues of the day, and people self-consciously interested in being a part of a public. Jean-Jacques Rousseau, the philosopher, and Jacques Necker, the French finance minister, were setting out their philosophical ideas justifying public opinion and its relationship to governance as well as coining the phrase *public opinion*.

The true origins of public opinion as we understand it today are more properly attributed to 17th-century England than to 18th-century France. England was a country profoundly influenced by the Reformation, which set off a burst of creativity in the printing industry. This led to economic and political conditions that produced a wide range of printed products, such as pamphlets and early newspapers, for a wide swath of society. England had a vigorous popular press and an egalitarian pub culture in which people of many diverse classes, informed by the pamphlets and early newspapers, met and discussed issues of the day. Framing the contemporary issues and providing concrete avenues for political activity was a vibrant system of petitions. These petitions, which were inexpensively produced and reproduced in large numbers by the commercial printers of the day, were debated in pubs across the country. Signed petitions representing the outcomes of these conversations were presented at court. Kings and their advisers for their own reasons felt duty-bound to reply to these petitions. This was not the elite salon culture of 18th-century France, but a broad-based social movement. English culture lacked the eloquence of Rousseau's philosophical arguments and the concise elegance of the French phrase *l'opinion publique*. But English society developed a working system of public opinion, based on a system of news and popular information, about a century before the French.

The American Revolution was influenced indirectly and directly by the experiences of both England and France. The American experiment in democracy settled on representative democracy under a constitution that guaranteed and safeguarded fundamental human rights of freedom of speech, press, petition, and religion. Also recognized were the rights of assembly and limited government power over individuals, establishing the principle of innocent until proven guilty to protect individuals from state power. The constitution also established a system of checks and balances among the various branches of government and procedures that empowered people but kept this power carefully controlled and subject to administrative and judicial approvals at various stages. This was the American solution to the issue of competence of public opinion. The republic was founded on the natural rights of people to self-govern, but governance itself was accomplished through layers of representatives under a constitution that protected fundamental liberties and human rights.

What Is Public Opinion?

Public opinion deals with opinions that people are willing to share with others, in contrast to their private opinions. Thus, communication emerges as a central variable in public opinion. Opinions formed but not shared are considered private or perhaps "secret opinions." Public opinion derives its force from people's willingness to express preferences "in public" and even to offer dissent from established policies. It is intimately related to issues of governance, and through public opinion people have asserted the right to self-government. This suggests that public opinion is primarily about "bottom-up communication" from citizens to their government. To the extent that government prevents the flow of such information or does not allow these claims, public opinion does not exist. The historian Richard J. Evans argued in his book *The Third Reich in Power* that public opinion was effectively extinguished in Germany in the 1930s by a series of repressive measures aimed at the news media, rival political parties, and even private conversation by Hitler's government. This suggests that public opinion cannot be just a collection of individual, privately held opinions.

However, if it is not just the aggregation of individual private opinions, what else can it be? Rousseau wrote of a concept he called "the General Will" or will of the nation. Some understand this as public opinion, but they are not synonymous. General Will is a highly abstract and mystical notion of a kind of supra-individual "mind of the nation" supposedly based on fully informed, rational thinking in the public's best interest. This concept was used to justify repression of dissenters during the French Revolution, and a century later arguably formed the basis for totalitarian regimes' control of people against their individual will. Rousseau himself contrasted the General Will with a more mathematical concept of a majority vote of the people, that is, as a result of an imaginary election in which individuals participated.

Gabriel Tarde, a criminologist, was also an early social theorist who was careful to explicitly link the emerging mass medium of newspapers to interpersonal communication and the development of the idea of a rational public. Tarde was profoundly concerned with both mass communication as able to channel what he called "currents of opinions" through the literate population without mass meetings and crowds intervening. Tarde noted the conditions for a "public" as being a group of readers—not just scholars—dispersed across a wide territory who were joined together by their shared thoughts on certain topics that were created by their reading of the same news reports. Tarde thus inextricably linked the public and the institution of journalism.

In the early 1990s, pioneering social psychologist Charles Horton Cooley defined *public opinion* not as the "mere aggregation of opinion" but as a "social product," that is, a result of communication and reciprocal influence. Cooley wrote an extensive treatise on his theory of public opinion in which he made clear that it was the product of coordinated communication in which people give time and attention to studying a problem individually. A nation can be said to have a form of public opinion when individuals must take into consideration their own ideas but also fresh ideas that flow from others engaged in thinking about the same topics. The outcome of the process for Cooley was not agreement but "maturity and stability of thought" that resulted from sustained attention and discussion about a topic. In other words, people might not agree after deliberation more than they did before, but their differences after careful consideration are better thought out and less changeable. Cooley believed that after such deliberation, people would know what they really think about a topic as well as what others around them think. To Cooley, this was the important difference between true opinion and simply one's casual impression. Today, we would refer to this as "opinion quality."

James Bryce was an early theorist of public opinion whose views, from the late 1880s, are of particular importance because of their influence on George Gallup, who did more than anyone in his era to bring together survey research, public opinion, and democracy. Bryce theorized that the history of democracy was best explained by stages in which political regimes took public opinion seriously. The first stage reflects systems in which public opinion is theoretically acknowledged as the basis of governance but the public is passive and allows a dominant group to exercise authority. The second stage moves theory into practice, such as when the early United States adopted a constitution built on the principle of the consent of the governed and the people begin to exercise power. The third stage reflects public opinion through regular, periodic elections in which people can shape government's direction by their votes. But Bryce acknowledged that voting was an imprecise way of forming public opinion and communicating it to the government. The fourth stage, according to Bryce, involved continuous monitoring of public opinion without the filtering of representatives or voting. The result would be popular sovereignty that no longer needed voting. Bryce's fourth stage was purely theoretical, however, largely because there was no practical way to achieve such continuous monitoring, and at the time he was writing he was doubtful such a process could ever be carried out in practice.

Gallup believed survey research and the resulting science of public polling were the tools to bring Bryce's understanding of the ultimate stage of democracy into existence, although he was careful to say that the purpose of polls was not to sabotage representative government but rather to provide "technical assistance" that would supplement and enhance the work of representatives. As Gallup explained, accurate public opinion polls would replace informal soundings of public preferences with a systematic, scientific set of procedures for knowing what people's opinions and preferences were. Thus, it is hard to overestimate Gallup's influence on contemporary thinking about public opinion, how it is measured, and its intimate relationship to survey research in contemporary life. Although there are many research tools available to provide information about public opinion, survey research is by far the most dominant, and Gallup was a leading figure in the development of the survey research industry.

Problems of Public Opinion and Its Measurement

Contemporary democracies measure their successes in various ways, chief among them on the nature and breadth of the population that constitutes the polity. Other measures include the equal treatment of all members of the society, the level of effective consultation or voice that citizens have influencing government actions, and the degree of legal protection citizens have from governmental intrusions or abuses. All of these criteria have something to do with public opinion. In fact, each democratic attribute raises significant questions about public opinion, at least indirectly. Some key questions in the study of public opinion through the ages are presented in the following subsections that directly or indirectly relate to the nature of democratic systems.

Who Is "the Public" and How Representative Are Opinion Polls?

Does the electorate include men and women? Are minorities included proportional to their numbers in the population? Although at the time of the American Revolution the public was defined narrowly to exclude women and African American slaves, full citizenship rights were eventually extended to all citizens and the right to vote for all guaranteed. However, even nowadays not all people are equally likely to respond to public opinion polls and surveys (e.g., opinion polls typically have differential nonresponse among the less educated strata of society), which therefore begs the question of how well opinion polls can measure public opinion in a society.

How Competent Are the People to Live Up to the Requirements of Democratic Systems?

This is perhaps the oldest and most important question about public opinion, and it has been asked since ancient times. Some of the early social scientists took a very dim view of the competency of public opinion. Gustav LeBon, one of the early writers of mass psychology, was a famous skeptic who took a dim view of "popular capabilities." Walter Lippmann was skeptical of the capabilities of public opinion. His influential books, *Public Opinion* and *The Phantom Public*, make the point clearly that democratic theory requires too much of ordinary people. Furthermore, the press, in Lippmann's view, was not adequate to the task of public enlightenment as required in democratic society.

When the issue of public competence is raised, it is useful to determine whether the supposed incompetence is attributed to inherent limitations of people, or is it more properly limited to the public's level of literacy and educational attainment? Other possibilities include the type and amount of communication that is available to the masses, the role models and values conveyed to the public about expectations related to civic virtues, whether the public actually has the requisite skills such as how to participate in dialogue, debate, and deliberation of political choices. Lippmann's chief critic was the philosopher and educational reformer John Dewey. His 1927 book, *The Public and Its Problems*, argued that the public was fully capable but that too many people lacked the required levels of communication and the quality of public media needed for informed decision making.

Do Majorities Exercise Tyranny Over Minorities?

Tyranny of the majority is a third enduring problem of public opinion. The fear is that people whose views are in the majority will use their power to oppress those holding minority views. This raises questions of civil liberties such as the protections of freedom of expression and assembly as well as protections for speakers from retribution. There has emerged in the literature of public opinion over the years considerable attention to social pressures that hinder people's willingness to speak their mind on controversial issues. These constraints range from people's unwillingness to be out of step with others to real fear of repercussions or reprisals from employers or government officials.

Are People Too Gullible in the Face of Persuasion?

Susceptibility to persuasion has emerged as a key issue for public opinion in the media age, since if the mass public is too gullible in the face of propaganda, the competence of the masses cannot be assured. Much of the vast literature on media effects is at least to some extent guided by this concern. In addition to the standard mass persuasion literature, newer work on framing raises the issue in a new way. Framing effects have been shown to be very powerful in influencing people's reactions to issues and political figures, although the vast majority of framing effects studies are performed in laboratories and use stimuli that are carefully balanced in ways that natural political discourse is not.

Opinion Quality and Public Opinion Measurement

One of the newest questions in contemporary public opinion research harkens back to the oldest literature on the role of conversation and deliberation in helping to create fully informed citizens with well-thought-out, stable opinions. To address questions of opinion quality, a number of measurement approaches have been advanced in recent years, ranging from the Choice Questionnaire to Deliberative Polls and to sophisticated experiments incorporating opportunities for people to discuss and deliberate before having their opinions measured.

Choice Questionnaire

Many issues facing policymakers involve complex scientific or policy information or difficult concepts about which the public has not had the opportunity to learn. In such circumstances, standard polling will provide unsatisfactory results, since only a tiny portion will have opinions and most will say they have never thought of the questions. Long sets of questions about which people have no experience will not be tolerated well by most survey respondents. The Choice Questionnaire is a type of survey that attempts to incorporate opportunities for the public to learn about the issue being surveyed before questions are asked. This must be done in a manner sensitive to the nature of the framing of the alternatives so that that the information presented does not result in decision bias.

Deliberative Polls

As described by James F. Fishkin, deliberative polls are a type of research in which randomly selected individuals are invited to participate in in-person events during which they will be presented with various oral and written briefing materials about topics to be considered. Participants are surveyed about their views on the topics when they are sampled and again at the end of the process. After considering the information presented, the participants are divided into groups for discussion and deliberation about the topics. No control groups are typically specified for such gatherings, but some investigators have effectively utilized control groups to be able to investigate the effects from the briefing materials separately from the effects of deliberation. As discussed extensively in the vast literature of deliberative democracy, deliberation is generally believed to have beneficial effects on quality opinion formation because opinions formed under conditions of deliberation will presumably have been carefully considered, be based on considerations of relevant facts, and have resisted challenges by others. This is thought to lead to more stable opinions that are more resistant to change than opinions formed in less rigorous circumstances.

Gerald M. Kosicki

See also Attitude; Attitude Stability; Deliberative Poll; Differential Nonresponse; Issue Definition (Framing); Gallup, George; Opinions; Poll; Public Opinion Research

Further Readings

Baker, K. M. (1990). *Inventing the French Revolution*. Cambridge, UK: Cambridge University Press.

Barone, M. (2007). *Our first revolution: The remarkable British upheaval that inspired America's founding fathers.* New York: Crown.

Bishop, G. F. (2005). *The illusion of public opinion*. Boulder, CO: Rowman & Littlefield.

Bryce, J. (1888/1995). *The American commonwealth*. Indianapolis, IN: Liberty Fund.

Cooley, C. H. (1909/2005). *Social organization*. New Brunswick, NJ: Transaction.

Dewey, J. (1927). *The public and its problems*. Athens, OH: Swallow Press.

Evans, R. J. (2005). *The Third Reich in power.* New York: Penguin.

Fishkin, J. S. (1995). *The voice of the people: Public opinion and democracy*. New Haven, CT: Yale University Press.

Gallup, G., & Rae, S. F. (1940). *The pulse of democracy: The public opinion poll and how it works*. New York: Simon & Schuster.

LeBon, G. (1896/1982). *The crowd: A study of the popular mind*. Marietta, GA: Cherokee.

Lippmann, W. (1922). *Public opinion*. New York: Free Press.

Lippmann, W. (1927/2004). *The phantom public*. New Brunswick, NJ: Transaction.

Neijens, P. (1987). *The choice questionnaire: Design and evaluation of an instrument for collecting informed opinions of a population*. Amsterdam: Free University Press.

Price, V., & Neijens, P. (1997). Opinion quality in public opinion research. *International Journal of Public Opinion Research*, *9*, 336–360.

Rousseau, J. J. (1762/1997). *The social contract and other later political writings*. Edited by V. Gourevitch. Cambridge, UK: Cambridge University Press.

Sniderman, P. M., & Theriault, S. M. (2004). The structure of political argument and the logic of issue framing. In W. E. Saris & P. M. Sniderman (Eds.), *Studies in public opinion: Attitudes, nonattitudes, measurement error and change* (pp. 93–132). Princeton, NJ: Princeton University Press.

Tarde, G. (1901/1969). The public and the crowd. In T. N. Clark (Ed.), *On communication and social influence: Selected papers* (pp. 277–294). Chicago: University of Chicago Press.

Weissberg, R. (2002). *Polling, policy and public opinion: The case against heeding the "voice of the people."* New York: Palgrave Macmillan.
Zaret, D. (2000). *Origins of democratic culture: Printing, petitions and the public sphere in early modern England.* Princeton, NJ: Princeton University Press.

PUBLIC OPINION QUARTERLY (POQ)

Public Opinion Quarterly (POQ) is the journal of the American Association for Public Opinion Research (AAPOR). Founded in 1937 at Princeton University, the journal is an interdisciplinary publication with contributors from a variety of academic disciplines, government agencies, and commercial organizations. The mission of the journal is to advance the study of public opinion and survey research. At this writing, POQ is the premier journal in this field, with high citation rankings in political science, interdisciplinary social science, and communication.

POQ was originally focused on the impact of the mass media on public opinion and behavior. The conjunction of two developments contributed to the journal's founding: the consolidation of channels of mass communication (at that time, print, motion pictures, and radio) and the growth of social science research on human behavior. The mid-1930s was a time of intense concern about the impact of messages distributed simultaneously to mass audiences. Franklin Roosevelt's "fireside chats," Father Coughlin's radio commentaries, and Adolf Hitler's speeches from Nuremburg were exemplars of the new phenomenon of a single persuasive voice reaching millions of people across the spectrum of society at the same time. This era also saw the consolidation of commercial interest in the opportunities presented by the mass media to influence public views and consumer behavior. Advertising and public relations had become significant societal institutions. By this time, too, it was standard for government entities to have press offices and information bureaus to attempt to affect public views of policies and programs. Public opinion polls, commercial ventures supported by media concerns, had become a significant focus of attention in elections and in commercial research. The rationale for POQ, presented in the editorial foreword to the first issue, was that mass opinion had become a major societal force whose genesis, shifting trends, and effects needed careful scientific study.

Those interested in the study of public opinion included academics across a range of fields—political science, psychology, sociology—as well as researchers employed in government and commercial organizations. POQ was established as a meeting place for these varied interests. Although the journal was housed at Princeton University and edited by academics at several universities, articles by advertising executives, playwrights, government ministers, and employees of media organizations also were published. Papers were devoted to the study of both elite, opinion-shaping organizations as well as to the understanding of mass opinion. Articles on propaganda could be found alongside papers on new developments in public relations. Descriptions of recent poll findings, examination of polling methods, and commentary on the impact of polls were common features. Thus, the journal combined analysis and critique with how-to lessons on measuring and influencing public opinion. It blended theoretical arguments with applied investigation.

In subsequent decades, POQ's mix of content has remained broadly the same, though the relative size of subject categories has shifted. The societal and technological forces that stimulated interest in the study of public opinion in the 1930s were transformed over time, just as the scholarly world evolved in new directions. The early research focus on channels of mass communication and their impact on mass opinion and behavior gave way by the 1960s to more complex and nuanced conceptualizations of public opinion formation and change. The journal tracked this evolution in articles published by members of Paul Lazarsfeld's Columbia School, including authors such as Elihu Katz and Joseph Klapper. At the same time, the world of social science publishing expanded. Newer journals (e.g., *Journal of Communication, Journal of Advertising, Journal of Advertising Research, Journal of Broadcasting and Electronic Media, Journal of Consumer Research, Journal of Marketing Research, Journal of Public Relations Research, Communication Research, Political Psychology, Political Behavior*) began to publish submissions on mass media effects and public opinion that might once have appeared in POQ.

Articles examining survey methodology, though present in the journal from the first issue, gained greater prominence over time. In the 1930s and

1940s, the focus of many papers was on methods of interviewing, as polling and survey organizations transitioned from the use of mail questionnaires, discredited particularly by the failure of the *Literary Digest* Poll in 1936 to accurately predict the outcome of the presidential election. Interviewer rapport and the match between interviewers and respondents on demographic characteristics were investigated by such scholars as Daniel Katz and Paul Sheatsley. In addition, there was discussion of question design and sampling methods and on the general "failure of the polls" in the 1948 Dewey-Truman election. Methodological concerns beyond the commercial polls were evidenced in Philip Hauser's paper on the design of the 1950 Census and by Hugh Parry and Helen Crossley's classic 1950 examination of response validity in their Denver community survey.

The 1960s brought a focus on survey methods in studies of race that has continued to this day. Papers by Howard Schuman and others on race of interviewer effects were particularly notable. The 1970s and 1980s featured a blossoming of research on survey methods, including the earlier focus on interviewers plus many papers on question wording and order and a major push on mode of interview effects (telephone vs. face-to-face contacts).

The prominence of survey methodology in POQ now routinely incorporates all of the preceding themes, plus sampling, post-survey adjustment, and the always present and increasingly crucial issue of nonresponse. Work on mode of interview now concerns Internet surveys and cellular phones as well as the older landline telephone, face-to-face, and mail varieties. The design of self-administered questionnaires, whether paper or Web based, is undergoing a major examination. The survey contexts in which these methodological studies occur include a wide range of substantive concerns, certainly more than would be conjured up by the term *public opinion*. The increasing output of strong methodological studies has led POQ to expand the number of its pages by 25% and to add a fifth topical issue devoted to some special and important aspect of survey methodology. This issue, in turn, serves as the basis for an annual AAPOR methods workshop. Moving forward, the journal should keep pace with the demand for methodological education.

Peter V. Miller

See also American Association for Public Opinion Research (AAPOR)

Further Readings

Katz, E. (1957). The two-step flow of communication: An up-to-date report on a hypothesis. *Public Opinion Quarterly*, *21*, 61–78.

Klapper, J. T. (1957). What we know about the effects of mass communication: The brink of hope. *Public Opinion Quarterly*, *21*, 453–474.

Parry, H., & Crossley, H. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, *14*, 61–80.

Public Opinion Quarterly: http://poq.oxfordjournals.org/

Schuman, H., & Converse, J. M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, *35*, 44–68.

Public Opinion Research

Public opinion research is an applied social science activity that features prominently in various academic disciplines, including communication, political science, psychology, public policy, and sociology. Public opinion research often has as its goal the quantification and description of preferences of large numbers of people in carefully described geographic areas such as cities, states, or nations. Often the specific time frame under study is particularly important, and typically public opinion is assessed on a recurring basis by firms and organizations specializing in such tasks. Some of these involve media organizations that use the results of public opinion research as the basis for news stories or to comment upon political or social events, policies, and people.

Use of Survey Research to Assess Public Opinion

Most often, public opinion as the object of empirical research is studied using some variation on the methods of survey research. A great deal of emphasis is typically placed on estimating the proportion of the adult or registered voter population holding certain attitudes toward various public affairs topics or evaluations of various public officials. Breakdowns by political party affiliation and various demographics are commonly presented. The accuracy of political polling is often verified by comparisons of poll results to election results. Although that works in an acceptable manner for polls that measure voter intention and are taken close to Election Day, it is less helpful for standard measures of political attitudes, for which there are no real-world indicators.

Much public opinion research is descriptive, that is, it attempts to estimate population parameters of various benchmark indicators such as presidential approval, the state of consumer confidence, and general levels of trust and confidence in government, among other standard questions. Such research is often done on a periodic basis, and recent results are compared to the running time-series of results of similar surveys done by the same organization. An alternative is to round up similar questions taken at the same time by different organizations, but such results are often difficult to compare directly because of slight variations in question wording, sampling and weighting strategies, various field work considerations, and other types of "house effects" that subtly influence survey results. Because such characteristics of polls or surveys tend to be consistent within the organization that produced the poll, it is usually considered most appropriate to compare results at a given time with previous results of the same poll.

Criticisms and Controversies

Public opinion research can be criticized in various ways, including that many results are less than fully useful because people have not considered the questions carefully or do not have well-formed opinions on particular matters and so invent an answer on the fly in response to a survey question. For years this sparked controversy among scholars and users of public opinion polling as the "pseudo-opinion" problem. More recent scholarly work that is informed by cognitive models addresses this by noting that such results may not be quite as serious of a problem as previously believed, largely because people may invent opinions on the spot but they do so using their values, social positions, and demographic background factors that are relatively stable. According to these "online processing models," such responses may have substantial validity even though people have not previously considered the issue.

Another criticism of contemporary public opinion research comes from Daniel Yankelovich, a leading pollster, social critic, and advocate of deliberative democracy who is chairman and cofounder of Public Agenda. In his 1991 book, *Coming to Public Judgment*, Yankelovich said that there is a missing concept in the way surveys currently are used to measure public opinion. What is missing is a way to measure the "quality" of the opinions that people express in surveys. To Yankelovich, high-quality public opinion is stable, consistent, and for which a respondent expressing the opinions recognizes their implications and takes responsibility for them. Since Yankelovich's critique, a number of scholars have conducted research to try to create ways to assess the quality of the opinions expressed by respondents in opinion polls, but to date no best practice has been agreed upon for doing this.

Various technical considerations of public opinion assessment have been controversial in recent years. Telephone polling became the dominant method of opinion assessment by the mid-1980s, replacing in-person interviewing, when landline telephone penetration became nearly universal. Random-digit dialing (RDD) technology combined with computerized programs that automate dialing and many other aspects of survey questionnaire administration revolutionized public opinion assessment and drove costs lower. However, other technological trends have continued to chip away at the dominance of telephone surveys and presented new challenges to methodologists, including the widespread use of automatic telephone answering machines, privacy managers, and caller identification systems, and other ways to screen telephone calls have made gaining survey responses by telephone more difficult and expensive. Patterns of assigning telephone numbers to electronic devices such as fax machines and computers have also made sampling populations by RDD more difficult. The widespread availability of wireless telephone technology and the migration of landline numbers to wireless devices have added complications. As residential consumers have switched from landline telephones to mobile ones, concerns about the adequacy of standard RDD sampling frames have increased. While the survey industry is developing reasonable solutions to these problems as they have appeared, it is not clear that public understanding has kept up. This may cause problems for the credibility of public opinion research among some consumers.

Other Ways to Assess Public Opinion

Public opinion research is nearly synonymous with *survey research*, but the two are not identical. In fact, public opinion can be assessed in many other ways besides survey research, and there are other schools of thought about the problem of public opinion. There are a number of alternatives to survey research, including qualitative interviews such as focus groups, content analysis, historical methods, and mass observation. None has the widespread use, flexibility, or quantitative nature of survey research. However, a popular public opinion textbook argues that survey research alone is an insufficient way to study public opinion and recommends a series of tools including content analysis, focus groups, and controlled laboratory experiments. In fact, considering the whole of public opinion, scholars from diverse disciplines use a wide range of evidence to make claims about public opinion, including focus groups, systematic analysis of mass media content, and other, more exotic tools as described following.

Experiments

In the past several decades, public opinion research has been enriched theoretically by the infusion of cognitive theory. Along with this has come increased interest in using the power of controlled laboratory experiments to help understand the psychological processes by which opinions are formed. Some of the best of these are careful about recruiting more representative samples than often is the case with experimental designs (e.g., subjects from the citizenry at large), but many rely on convenience samples of college students. Experimental studies tend to be particularly useful for understanding the processes by which people incorporate new information, such as media messages, and for building cognitive theory. Moving ideas back and forth between experiments and survey research seems to be a key characteristic of the growing subfield known as *political psychology*. Another use of experimental logic and design also has increased dramatically in recent years due to the advent of computerized data collection systems that make it relatively easy to provide alternative question forms to randomly selected subgroups within the main survey sample. This has facilitated a burgeoning interest in question-wording experiments and has hastened the acceptance of cognitive models into the survey research literature.

Focus Groups

Focus groups are a type of qualitative research tool that relies on moderated group interviews of 6 to 12 people at a time. Typically lasting an hour or two, focus groups are often transcribed or recorded on audio or video for further study. The experience is organized with a carefully planned conversation guide to move people through the various topics under consideration. Moderators must be able to create a nonthreatening tone for the conversation and encourage everyone to share their opinions while preventing talkative participants from dominating the conversation. Focus groups develop a unique group dynamic based on the individuals who participate in any given session, and this needs to be accounted for in any analysis of focus group data. Focus groups are used as aids in designing more systematic survey questionnaires, but they have often proved valuable in helping explain polling results and explore processes uncovered in survey research. The best focus groups typically use random sampling to select participants.

Content Analysis

Content analysis is a systematic, rigorous, and quantitative analysis of printed or visual materials. Because of the intimate connection between public opinion and popular media, many scholars believe that systematic analysis of media content helps provide clues about what material informed public opinion in a given time and place. Carroll J. Glynn and her coauthors argue that media content can reveal valuable evidence about public opinion because their popularity indicates that they “resonate with cultural norms, values and sentiments.” For example, Media Tenor, a leading global firm based in Germany, specializes in content analysis of the world’s media for a wide range of corporate and government clients.

Historical Methods

There is intense interest in understanding public opinion at key points in history. For historical periods for which there is survey research available, the problems are mainly to locate the original raw data and translate it in such a way that it can be analyzed by contemporary computers and software. Early survey data were produced using certain data coding conventions that are no longer used, including multi-punched data that were designed to accommodate more than 10 symbols in a single digit space. This was accomplished by adding extra, nonnumerical symbols such as @, &, or # to the data field. In an era when data were entered onto computer cards and sorted by machine, this would have helped reduce sorting and thus provided increased efficiencies in data processing, and such systems were used by some polling firms at least into the 1980s. Other issues with reanalyzing older data relate to the nature of the quota and other nonprobability sampling procedures that were popular at the time. This may require special weighting and analysis procedures. A surprising amount of data from earlier times survives and is available for reanalysis at major data archives such as the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan, and the Roper Center for Public Opinion Research at the University of Connecticut. A major data source for Europe and the United Kingdom is the UK Data Archive at the University of Essex.

For time periods that pre-date the availability of survey data, historians have become adept at finding other types of data with which to pursue studies of public opinion. Typical source material includes diaries and records, political petitions, legislative records, letters, and other materials limited only by imagination and availability. Robert Darnton, a leading historian of the French Revolution, for example, has uncovered a rich trove of records of expressed opinion in France on the eve of the French Revolution in the basements of the Bastille. These investigative records were created by secret police employed by Louis XV as he tried to hold on to power in the final years leading up the revolution. They include records of overheard conversations in public places, restaurants, and other settings where people gathered to discuss public affairs. There is a fast-growing historical literature examining public opinion at various key points in history by resorting to a wide variety of data sources in order to make inferences. For example, David Zaret has made a career examining the use of popular petitions in England in the 17th century to understand popular resentments and problems that were brought to the attention of monarchs. One advantage of these petitions as raw data for the analysis of public opinion is that they were signed by people of all social classes and in many cases were acted upon favorably by the king. Public opinion during the American Civil War is another example. Much of the professional work of popular historian Stephen Ambrose on World War II soldiers was conducted via oral histories of the men and women who served in the American military during the war.

Mass Observation

Emphasizing the study of expressed opinion—that is, opinions that people shared with one another in public settings—was a major goal of an organization known as Mass Observation that was active in Britain starting in the 1930s. The group used anthropological-style field methods to observe people in conversations in public places, measuring what opinions people expressed in public about topics of concern. The group was founded by journalist Charles Madge, poet and filmmaker Humphrey Jennings, and self-taught anthropologist Tom Harrisson. The three men envisioned Mass Observation as a scientific organization for the assessment of public life. Over the years it grew into a type of social movement, at its high point involving thousands of volunteers. The organization benefited from contracts from the British government during World War II to gather systematic evidence about morale and the life of citizens as impacted by war and continued to exist after the war. The movement eventually became a private firm engaging mainly in consumer market research and eventually merged with BMRB International, a large British advertising agency. Mass Observation continues to exist to this day as a data archive at the University of Sussex. The Sussex group continues to solicit participants in its panel of writers. In recent years a number of reports from the original Mass Observation project have been reissued, and several new ones compiled. Unlike survey research, which gathers self-reports of individuals' private opinions expressed only to an interviewer, Mass Observation gathered the comments and opinions that people actually expressed in public settings.

Gerald M. Kosicki

See also Approval Ratings; Consumer Sentiment Index; Content Analysis; Convenience Sampling; Experimental Design; Focus Group; Media Polls; Nonattitude; Precision Journalism; Public Opinion; Random-Digit Dialing (RDD); Roper Center for Public Opinion Research; Trust in Government

Further Readings

Berinsky, A. J. (2006). American public opinion in the 1930s and 1940s: The analysis of quota-controlled sample survey data. *Public Opinion Quarterly*, *70*, 499–529.

Calder, A. (1985). Mass observation 1937–1949. In M. Bulmer (Ed.), *Essays on the history of British* *sociological research* (pp. 121–136). Cambridge, UK: Cambridge University Press.

Campbell, D. A. (2003). *English public opinion and the American Civil War*. London: Royal Historical Society Boydell Press.

Darnton, R. (2004). Madamoiselle Bonafon and the private life of Louis XV: Communication circuits in eighteenth-century France. *Representations*, *87*, 102–124.

Glynn, C. J., Herbst, S., O'Keefe, G. J., Shapiro, R. J., & Linderman, M. (2004). *Public opinion* (2nd ed.). Boulder, CO: Westview.

Herbst, S. (1993). *Numbered voices: How opinion polling has shaped American politics.* Chicago: University of Chicago Press.

Hubble, N. (2006). *Mass-observation and everyday life: Culture, history, theory*. London: Palgrave Macmillan.

Iyengar, S., & McGuire, W. J. (Eds.). (1993). *Explorations in political psychology.* Durham, NC: Duke University Press.

Sirkin, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Tourangeau, R. (1999). *Cognition and survey research*. New York: Wiley.

Yankelovich, D. (1991). *Coming to public judgment: Making democracy work in a complex world.* Syracuse, NY: Syracuse University Press.

Zaret, D. (2000). *Origins of democratic culture: Printing, petitions and the public sphere in early-modern England.* Princeton, NJ: Princeton University Press.

Purposive Sample

A purposive sample, also referred to as a *judgmental* or *expert sample*, is a type of nonprobability sample. The main objective of a purposive sample is to produce a sample that can be logically assumed to be representative of the population. This is often accomplished by applying expert knowledge of the population to select in a nonrandom manner a sample of elements that represents a cross-section of the population.

In probability sampling, each element in the population has a known nonzero chance of being selected through the use of a random selection procedure. In contrast, nonprobability sampling does not involve known nonzero probabilities of selection. Rather, subjective methods are used to decide which elements should be included in the sample. In nonprobability sampling, the population may not be well defined. Nonprobability sampling is often divided into three categories: purposive sampling, convenience sampling, and quota sampling.

An example of purposive sampling would be the selection of a sample of universities in the United States that represent a cross-section of U.S. universities, using expert knowledge of the population first to decide with characteristics are important to be represented in the sample and then to identify a sample of universities that meet the various characteristics that are viewed as being most important. For example, this might involve selecting large, medium, and small public and private universities. This process is referred to as *two-stage sampling*, but the first-stage units are not selected using probability sampling techniques. This kind of two-stage sampling should not be confused with *stratified sampling*, which is a probability method of sampling. Instead, in this example of purposive sampling, such "strata" (size of university and type of university) are used without the scientific rigor of probability sampling. Nonetheless, they are used to assure a more representative mix of elements than may otherwise occur if the expert were not explicitly choosing universities of different sizes and types.

Another example of purposive sampling could be the selection of a sample of food stamp offices from which participants will be sampled. Here, the first-stage units are selected to represent key food stamp recipient dimensions, with expert subject matter judgment used to select the specific food stamp offices that are included in the study.

One limitation of purposive sampling is that another expert would likely come up with different sampled elements from the target population in terms of important characteristics and typical elements to be in the sample. Given the subjectivity of the selection mechanism, purposive sampling is generally considered most appropriate for the selection of small samples often from a limited geographic area or from a restricted population definition, when inference to the population is not the highest priority.

Michael P. Battaglia

See also Convenience Sampling; Nonprobability Sample; Probability Sample; Quota Sampling; Stratified Sampling

Further Readings

Henry, G. (1990). *Practical sampling.* Newbury Park, CA: Sage.

Push Polls

A push poll is a form of negative persuasion telephone calling during a political campaign that is meant to simulate a poll but is really intended to convince voters to switch candidates or to dissuade them from going to the polls to vote. To an unskilled recipient of such a call, it sounds like a traditional telephone survey at the start, but the tone and content of the questioning soon changes to the provision of negative information about one of the candidates. A distinguishing characteristic of a push poll is that often none of the "data" are analyzed; the purpose of the call is to "move" voters away from a preferred candidate. Push polling is so antithetical to legitimate polling that in 1996 the American Association for Public Opinion Research (AAPOR), the American Association of Political Consultants (AAPC), and the National Council on Public Polls (NCPP) issued a joint statement condemning the practice. Since then, the Council for Marketing and Opinion Research (CMOR) has joined in this denunciation.

The support of the AAPC was important because campaigns often test thematic content or different approaches to arguments for the purposes of developing campaign strategy. In these circumstances, the polling firm is collecting data for analysis to determine which themes or approaches are most effective. But in a push poll, no data are analyzed; the calls are made with the intent of getting voters to switch their support from one candidate to another or to so turn them off from the political process that they will decide to stay home on Election Day rather than vote. And a great many more people are contacted than are needed in the standard sample size for a legitimate election campaign poll.

In a period of declining response rates, polling and survey research organizations are always concerned about establishing good rapport with respondents, in the short term with an eye toward completing a specific interview and, more generally, with the purpose of maintaining a positive image of the industry that will promote future cooperation in surveys. Having a bad experience with something that seems like a very biased poll is harmful to both of these interests.

In general, push polling efforts occur late in the campaign, when public disclosure is more problematical. Many of those called cannot distinguish between a legitimate call from an established polling firm for a telephone interview and a push poll call, so they stay on the phone. If the push poll campaign is not brought to light quickly, Election Day arrives and news coverage wanes right after a winner is declared in the contest; so the push poll succeeds in its intent without disclosure. Because there is little regulation of the polling industry, and none in real time, the most effective antidote to push polling is disclosure through the media. But this requires learning of the push polling through those who have received the calls and then trying to track down the source of the calls. Those who are most sensitive to push polls and likely to be sources of disclosure are pollsters and campaign workers. However, any single person is very unlikely to receive such a call to begin with.

Push poll calls are typically very short and involve only a few questions, unlike the typical telephone interview that lasts at least 5 minutes and sometimes up to half an hour. The questions often involve significant distortions of one candidate's record or positions. Push polls sometimes involve tens of thousands of calls being made in a state where there is a presidential primary taking place, or just a few thousand calls for a smaller constituency. Of course, no individual who receives such a call can tell from their own experience that they are part of a larger campaign, making detection of the negative calling campaign more difficult.

In their original form, these calls were made by interviewers who worked from a script. With the advances of technology, they can now be produced on a computer that does both the dialing and reading of the script. This means that more calls can be made in a shorter period of time and at a lower cost per call.

Michael W. Traugott

See also American Association for Public Opinion Research (AAPOR); Council for Marketing and Opinion Research (CMOR); National Council on Public Polls (NCPP); Poll; Pseudo-Polls

Further Readings

American Association for Public Opinion Research. (2007, June). *AAPOR statement on "push" polls.* Retrieved June 12, 2008, from http://www.aapor.org/aaporstatementonpushpolls

American Association of Political Consultants. (n.d.). *AAPC board of directors declaration regarding "push polling."*

Retrieved June 12, 2008, from http://www.theaapc.org/about/pushpolling/

Council for Marketing and Opinion Research. (n.d.). *What is research? Differences between research and sales.* Retrieved from http://www.cmor.org/research/differences.cfm

National Council on Public Polls. (1995, May 22). *Push polls*. Retrieved June 12, 2008, from http://www.ncpp.org/?q=node/41

p-VALUE

The probability value (abbreviated "*p*-value"), which can range from 0.0 to 1.0, refers to a numeric quantity computed for sample statistics within the context of hypothesis testing. More specifically, the *p*-value is the probability of observing a test statistic that is at least as extreme as the quantity computed using the sample data. The "probability" is computed using a reference distribution that is generally derived using the null hypothesis. The phrase *extreme* is usually interpreted with respect to the direction of the alternative hypothesis—if two tailed, then the *p*-value represents the probability of observing a test statistic that is at least as large in absolute value as the number computed from sample data. Put another way, the *p*-value can be interpreted as the likelihood that the statistical result was obtained by chance alone.

P-values often are reported in the literature for analysis of variance (ANOVA), regression and correlation coefficients, among other statistical techniques. The smaller the *p*-value, the more the evidence provided against the null hypothesis. Generally, if the *p*-value is less than the level of significance for the test (i.e., α), then the null hypothesis is rejected in favor of the alternative and the result is said to be "statistically significant."

In the context of survey research, interest may be given in comparing the results of a battery of questionnaire items across demographic or other groups. For example, a sequence of *t*-tests may be conducted to compare differences in the mean ages, education level, number of children in the home, annual income, and so on, for people who are surveyed from rural areas versus those from urban areas. Each test will generate a test statistic and a corresponding two-sided *p*-value. Comparing each of these separate *p*-values to .05 or .10, or some other fixed level of alpha, can give some indication for the significance of the hypotheses tests that the mean levels of age, education, and so on, are equal for citizens living in urban and rural areas versus the alternative hypothesis that these means are different. However, using the separate *p*-values alone inflates the Type I error rate of the entire sequence of tests. To avoid this inflation, adjusted *p*-values are often used to provide an overall error rate that is equal to the nominal alpha level specified by the researcher. The most straightforward of these adjustments is the Bonferroni adjustment, which compares either the *p*-value to alpha divided by the number of tests or compares the *p*-value multiplied by the number of tests to the nominal alpha level. If the adjusted *p*-values are less than alpha, then the specific null hypothesis is rejected.

The computation of *p*-value can be illustrated using data obtained from a single survey item that asks a random sample of 20 households the following question: *How many children live full time within your household?* The null hypothesis is that an average of 2 children live full time per household; the alternative hypothesis posits the average number exceeds 2. The reference distribution for the sample mean values for samples of size 20 is provided in Table 1.

The sample mean computed from the 20 sampled households was 2.5 children per household. In this case the hypothesis is one-tailed and the *p*-value is literally interpreted as the probability that the sample

Table 1 Distribution of the sample mean number of children per household for possible samples of 20 households randomly selected from the population

Possible Values for $\bar{X}$	*Proportion of Samples Having This Value for* $\bar{X}$
0	5%
1	10%
1.5	15%
2	35%
2.5	20%
3	10%
3.5	5%

mean is 2.5 or larger. To compute this probability, we use the reference distribution from Table 1 as follows:

$$\begin{aligned} p\text{-}value &= \text{Probability } (\bar{X} \geq 2.5) \\ &= \text{Probability } (\bar{X} = 2.5, 3 \textit{ or } 3.5) \\ &= .20 + .10 + .05 = .35. \end{aligned}$$

In this case, the result is interpreted as not statistically significant since the *p*-value was .35, which exceeds nominal alpha levels of .05 or .10.

Trent D. Buskirk

See also Alpha, Significance Level of Test; Alternative Hypothesis; Null Hypothesis; *p*-Value; Statistic; Type I Error

Further Readings

Le, C. T. (2001). *Health and numbers* (2nd ed.). New York: Wiley-Liss.

Yates, D., Moore, D., & McCabe, G. (1999). *The practice of statistics.* New York: W. H. Freeman.

Q

QUALITY CONTROL

The term *quality control* refers to the efforts and procedures that survey researchers put in place to ensure the quality and accuracy of data being collected using the methodologies chosen for a particular study. Quality-control efforts vary from study to study and can be applied to questionnaires and the computerized programs that control them, sample management systems to ensure proper case processing, the monitoring of appropriate interviewer behavior, and other quality-control aspects of the survey process, all of which can affect the quality of data and thus the results.

Training Interviewers

In surveys that use interviewer-administered questionnaires, proper training of interviewers is often seen as the heart of quality control. To become successful and effective, interviewers first must be introduced to the survey research process through a general training, in which fundamental concepts such as basic interviewing techniques, obtaining cooperation, and maintaining respondent confidentiality are covered. Following general training, interviewers should be trained on the specifics of an individual project. Training sessions should strive to maximize active participation. Topics covered should be reinforced with group discussion and interaction, trainer demonstrations, and classroom practice and discussion. Role playing and practice are also key elements of effective basic project training. Assessment of the knowledge gained and retained by interviewers is also a part of survey quality control.

Monitoring Interviews

In surveys that use interviewer-administered questionnaires, monitoring interviews has become a standard quality-control practice. Telephone interviewer monitoring is accomplished by using silent monitoring equipment so that neither the interviewer nor the respondent is aware of it. In addition, facilities may use equipment that allows the monitors to listen to an interview and simultaneously visually observe the progress of an interview on a computer screen. Monitoring serves four objectives, all aimed at maintaining a high level of data quality. These objectives include (1) obtaining information about the interview process (such as procedures or interview items) that can be used to improve the survey, (2) providing information about the overall data quality for the survey in order to keep the data collection process in statistical control, (3) improving interviewer performance by reinforcing good interviewing behavior, and (4) detecting and preventing data falsification.

Generally speaking, there are two different types of telephone monitoring conducted: *quality assurance monitoring* and *interviewer performance monitoring*. Quality assurance monitoring involves the use of a coding scheme that allows monitors to score interviewers on a variety of aspects such as proper question delivery, use of nondirective probing, and correct entry of responses. Interviewer performance monitoring takes more of a qualitative approach. Under this process, monitors record any observations that might be useful in subsequent evaluation of the interviewing staff. Monitors will make observations that focus on appropriate administration of items, remaining neutral, tone of voice, pace, and so on, and will provide feedback, both in the form of positive reinforcement and constructive criticism with potential retraining as warranted.

While monitoring telephone interviewers is an efficient and nonintrusive activity, conducting monitoring of face-to-face interviews in the field is more challenging. Traditionally monitoring face-to-face interviews in the field is done in one of two ways. First, direct observations can be made by field supervisors who accompany field interviews during data collection. During these periods, supervisors will observe and evaluate field interviewers on adherence to survey protocols while noting areas requiring improvement. Second, tape recordings are used to record portions of interviews and are reviewed after the fact to evaluate interviewers on quality, proper administration of the survey, and so forth. While effective in capturing quality information, these two methods have traditionally come with a cost, as they are both intrusive and cumbersome. Recent advancements in digital recording have made monitoring possible without the need to carry recording equipment such as microphones and tapes. *Computer audio-recorded interviewing* allows for the recording of portions of interviews for quality control on computer-assisted personal interviewing (CAPI) laptops, eliminating the need for external equipment.

Supervisory Meetings

During field and telephone data periods, supervisors often hold regularly scheduled meetings, sometimes referred to as *quality circle meetings*, to discuss data collection issues. These sessions are conducted to build rapport and enthusiasm among interviewers and supervisory staff while assisting in the refinement of an instrument and providing ongoing training for the staff. Such meetings have the potential to identify previously unrecognized problems with an instrument, such as questions that respondents may not understand or questions that are difficult to administer. In addition these sessions may reveal computer-assisted interviewing software problems and issues with data collection procedures and operations.

Verification Interviews

In addition to monitoring for quality and conducting regular meetings with interviewers, many data collection supervisors take an additional step to ensure data quality and prevent falsification. Verification by telephone and reviewing interview timing data are particularly useful and important for field interviews where data are not collected in a central facility; when data are collected this way, it makes the interviewers more susceptible to falsification, also known as "curb stoning."

Telephone verification is conducted by first telling respondents who have just completed face-to-face interviews that they may be contacted at a later date for quality-control purposes. Telephone interviewers will then contact selected respondents to verify the face-to-face interview was completed properly and with the correct sample member. Any suspicious results are then conveyed to the data collection supervisors for investigation.

Some face-to-face surveys also review questionnaire timing data as part of standard quality-control procedures. Under this method, as interviews are completed and transmitted from the field, timing data are captured for each module or section within the instrument, along with overall timing data on the overall length of the study. By recording and analyzing this information, researchers can create reports that can be used to detect suspiciously short interview administration times. This quality-control method allows for early detection and resolution of problems, leading to more complete and accurate data for analysis.

Examining Systems and Data for Quality

In addition to training and implementing procedures to ensure quality control through interviewers, and regardless of the mode of data collection and whether or not the questionnaire is interviewer administered, researchers must also take steps to verify that the systems and procedures being used to collect and process data are functioning properly. By using systematic testing plans, researchers can verify routes through the computerized questionnaire and ensure that a program is working as specified in terms of display, skip pattern logic, and wording. Researchers should also review frequencies of the collected data throughout data collection (especially early in the field period) to ensure skip patterns are being implemented properly and to identify and address any questions that have higher-than-expected rates of item-level nonresponse. Logic checks should be deployed to evaluate the likely accuracy of the final data set before analyses begin. If coding of open-ended responses is conducted, quality-control measures must be deployed to ensure inter-coder reliability. Finally, reliability checks should be conducted on all analyses that are performed.

David James Roe

See also Coding; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Face-to-Face Interviewing; Falsification; Field Period; Frequency Distribution; Intercoder Reliability; Interviewer Monitoring; Interviewer Training; Missing Data; Supervisor; Validation; Verification

Further Readings

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality.* Hoboken, NJ: Wiley.

Fowler, F., & Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: Sage.

Lavrakas, P. (1993). *Telephone survey methods: Sampling, selection, and supervision.* Newbury Park, CA: Sage.

Quality of Life Indicators

Two types of survey measures fall under the quality of life domain. The first is *psychological measures* used to assess satisfaction in response to a life situation, typically an individual-level measure of quality of life in the context of the mental health of persons living with chronic health diseases. The second is a *social indicator*, designed to assess life quality for the purpose of creating aggregated indices for making systemic comparisons. Though the former has received a higher volume of research, the latter is more central to public opinion research.

Assessing the well-being of individuals and communities may be one of the most important contributions public opinion researchers can make. By providing benchmarks, researchers can elucidate the antecedents and consequences of individual life satisfaction as well as assess the progress of communities over time. At the same time, measuring quality of life is extremely complex as it encompasses myriad dimensions, including both material conditions and psychological orientations.

The challenge for social indicators researchers has been to come up with objective measures that are both comprehensive, in terms of capturing the various relevant dimensions of quality of life, and non-ethnocentric, in that they do not reflect researcher biases in terms of what constitutes life quality. However, any attempt to measure quality of life inherently involves making subjective decisions about what constitutes "quality." As such, these measures will always be subject to charges of ethnocentricity. However flawed such measures are, they do serve as a point of comparison and a yardstick against which to measure progress.

The unit of analysis is one important consideration in the assessment of life quality. That is, will the measure describe differences in the life quality of individuals, communities, or nations? At the micro level, measures attempt to assess the psychological and material well-being of individuals. Other indicators have been designed to assess the quality of life afforded by a particular community or nation. Macro-level indices most often are based on indicators that are aggregated and reported in terms of mean levels; however, when it comes to larger social units, the dispersion of such aggregated measures is also significant in order to take into account inequality in the distribution of resources that contribute to quality of life.

Quality of life measurement has typically involved a combination of economic, psychological, and social conditions. Economic indicators include estimates of standard of living, disposable income, wages, savings, bankruptcy rates, and other such criteria. Among the dimensions of social conditions are measures of crime, education, employment, crowding, pollution, aesthetic surroundings, housing quality, birth rates, infant mortality, longevity, doctors per capita, and health care. Psychological dimensions include assessments of happiness, life satisfaction, self-esteem, self-efficacy, marital satisfaction, family life, friendships, health, housing, jobs, housework, neighborhoods, communities, education, standard of living, financial state, and life in general. These measures may be classified into two groups: external measures, which refer to measures of the structural conditions that impinge upon an individual (often referred to as *social indicators*), and internal measures, which refer to indicators based on respondents' perceptions of their own life situations (often referred to as *measures of social well-being*).

Some prominent examples of quality of life measures are Cantril's Self-Anchoring Scale; Campbell, Converse, and Rogers's Domain Satisfactions and Index of Well-Being; Andrews and Withey's Global Well-Being Index; and Deiner's Quality of Life Index. Among the measurement challenges in this area are the following: the positivity bias (the tendency for respondents to use the high end of the scale for perceptual measures of psychological indicators), the ethnocentrism bias (indicators that reflect values of the researchers), the subjectivity bias (a reliance on nonobjective perceptual indicators), the quantitative bias (a reliance on quantitative measures that may be less meaningful than qualitative indicators), measurement biases (such as response set and social desirability), weighting biases (problems of determining the relative importance of indicators), and measurement complexity (because there are so many important dimensions of life quality, measurement is cumbersome).

Douglas M. McLeod

See also Bias; Level of Analysis; Perception Question; Positivity Bias; Public Opinion; Social Desirability

Further Readings

Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: American perceptions of life quality.* New York: Plenum Press.

Campbell, A., Converse, P. E., & Rodgers, W. L. (1976). *The quality of American life.* New York: Russell Sage Foundation.

Cantril, H. (1965). *The pattern of human concerns.* New Brunswick, NJ: Rutgers University Press.

Deiner, E. (1995). A value based index for measuring national quality of life. *Social Indicators Research*, *36*, 107–127.

QUESTIONNAIRE

The questionnaire is the main instrument for collecting data in survey research. Basically, it is a set of standardized questions, often called *items*, which follow a fixed scheme in order to collect individual data about one or more specific topics. Sometimes questionnaires are confused with interviews. In fact, the questionnaire involves a particular kind of interview—a formal contact, in which the conversation is governed by the wording and order of questions in the instrument. The questionnaire often is administered in a standardized fashion, that is, in the same way to all the respondents of the survey. The logic behind the standardization of questions and answers is that only if a stimulus is the same for all the respondents of a survey can it be, at least theoretically, possible to get the same (symbolic, cognitive, psychological, social) reaction from it. Responses obtained across individuals should be comparable to one another.

This entry discusses the construction, format, and layout of questionnaires, along with question wording and types. Lastly, this entry addresses the importance of pretesting and the various ways questionnaires can be administered.

Questionnaire Construction

When questionnaires are constructed, four primary requirements must be met:

1. Theoretical knowledge of the topic of research, achieved through the reconnaissance of the relevant literature (if such exists) and/or in-depth interviews or other qualitative methods of research (ethnographies, focus groups, brainstorming, etc.) that may serve as pilot studies.
2. Valid and reliable operationalization of concepts and hypotheses of research. Most questionnaire items, in fact, originate from the operationalization phase. To check the validity (the degree to which an item or scale measures what it was designed to measure) and reliability (the consistency or replicability of measurements) of a set of items, various techniques can be used: external, construct, and face validity, among others, in the first case; and parallel forms, test–retest, split-half, intercoder techniques, in the case of reliability.
3. Experience in writing a questionnaire, or at least the availability of good repertoires of published questionnaires.
4. A knowledge of the target population. This is crucial information: The target population must be able to answer to the questions accurately.

Questionnaires are usually composed of three main parts: the cover letter (or introduction), the instructions, and the main body. Usually, they finish with thanking the respondents for their valuable collaboration.

The *cover letter* (or its equivalent in interviewer-administered surveys) introduces the research and tries to motivate the respondents to cooperate with the survey task. It also explains the aims of the research, informs about its contractors and sponsors, and, above all, guarantees the anonymity or at least the confidentiality of the respondents. It is a sort of "contract," where the costs and benefits for collaboration between the respondent and the researcher are defined. The cover letter is one of the key elements in improving the response rate.

Instructions are especially important when the questionnaire is self-administered. Instructions contain all the rules the respondents must follow to answer the questions (e.g., how to check the boxes, which part of the questionnaire has to be skipped in certain cases, etc.). These rules should be as simple as possible. They can be categorized as (a) general instructions, (b) section introductions, (c) question instructions, and (d) "go to" instructions for contingency questions.

Finally, the *main body* includes the actual questions. There are many types of questions: questions about what people are (demographic data and attributes, such as gender, age, education, occupation), do (behaviors, such as buying records or traveling), think (beliefs, opinions, or judgments), or remember. These ingredients must be arranged in a structure that takes into account the attention, memory, sensibility, motivations, and background characteristics of the respondents.

Questionnaire Structure and Format

The structure of the questionnaire must be logical and the questions adequately linked and arranged together (e.g., grouping all the questions about the same subject together or using linking sentences when passing from one topic to another). Moreover, the length of the questionnaire must be reasonable: Only questions that are absolutely necessary should be included. With respect to questionnaire layout, at least these basic rules should be followed: (1) Each questionnaire must have an identification code, number, or both; (2) each question must have its own progressive number, and the space and graphic style of questions and response categories must be legible; and (3) the numerical values or codes for closed-ended questions (e.g., from the codebook) should be embedded into the questionnaire, in order to facilitate data entry into a case-by-variable matrix. In short, each response alternative must have its corresponding code (see examples in the following section of this entry). Of course, this does not apply to open-ended questions.

Generally, the ordering of items within a questionnaire follows this pattern: (a) general and neutral questions (to build rapport and thereby obtain the respondent's confidence), (b) questions that require greater effort (e.g., complex, core questions), (c) sensitive questions, and (d) demographic questions. This general pattern has been found to increase data quality for most surveys.

Nevertheless, the order in which questions are organized and the use of filter questions can yield various shapes, for example, funnel format (from general to specific questions), inverted funnel format (from specific to general questions), diamond format, X-format, box format, and mixed formats. The researcher will need to choose the form that best fits his or her goals. Classic and recent studies confirm that the order of the questions is an important contextual factor that influences respondents' answers.

Questionnaire Wording and Question Types

A general rule, in constructing questions, is to be *clear* in terminology and *simple* in structure. More specifically:

- Questions should use simple vocabulary.
- Their syntax should be simple, without subordinate clauses and without double negatives.
- They should not contain two questions in one (double-barreled questions).
- Questions must be concrete with respect to time and events.
- They should not lead the respondent to particular answers.
- The number of response alternatives should be limited unless additional visual cues are employed.
- All the alternatives of response should appear acceptable, even the most extreme.
- The response alternatives should be exhaustive and mutually exclusive.

Basically, two broad types of questions can be distinguished: *open-ended* and *closed-ended* (or *fixed alternative*) questions. The latter are more frequent in survey research. Open-ended questions are suitable when the researcher thinks it would be better to leave the respondents free to express their thoughts with their own words. However, they require a subsequent postcoding, for instance through content analysis, lexical correspondence analysis, or a qualitative treatment.

The difficulties and costs in the coding and analysis of open-ended questions have long limited their use in questionnaires. Today, given the huge advances in the field of (automatic) textual analysis, this limitation has lessened. On the other hand, closed-ended questions allow immediate statistical treatment. Unfortunately, sometimes the respondent cannot find a suitable answer among the alternatives proposed. In short, closed-ended questions require particular attention in providing relevant response alternatives.

There are many types of closed-ended questions. The following are examples of the most common:

- Selection among nominal categories, as in the following example:

  What is your religion?
  Christian [ ]$_1$
  Jewish [ ]$_2$
  Muslim [ ]$_3$
  Other [ ]$_4$
  Atheist/Agnostic [ ]$_5$

- Checklists, as in the following example:

  What kind of TV programs do you like watching? (Please check all the answers that apply.)
  News [ ]$_1$
  Sports [ ]$_2$
  Telefilms [ ]$_3$
  Soap operas [ ]$_4$
  Movies [ ]$_5$
  Talk shows [ ]$_6$
  None of the above [ ]$_7$

- Selection among ordinal categories, as in the following example:

  How many times did you go shopping last week?
  Three or more [ ]$_4$
  Twice [ ]$_3$
  Once [ ]$_2$
  Never [ ]$_1$

- A particular kind of selection among ordinal categories is the degree of agreement or disagreement with a statement, as in the next example, known as a Likert-type question:

  Please indicate how much you agree or disagree with the following statement: "I find much of my job repetitive and boring."
  Strongly Agree [ ]$_5$
  Agree [ ]$_4$
  Neither Agree nor Disagree [ ]$_3$
  Disagree [ ]$_2$
  Strongly Disagree [ ]$_1$

- Ranking of personal preferences, as in the following example:

  Please rank from 1 (most favorite) to 5 (least favorite) the following drinks, using each number only once:
  a. Wine _____
  b. Beer _____
  c. Cola _____
  d. Water _____
  e. Lemonade _____

- Semantic differential scaled responses, as in the following example:

  How happy are you with the government's choice of diminishing funds for military research?
  Very Happy 7…6…5…4…3…2…1 *Very Unhappy*

- Interval-level or ratio-level responses, as in the following example:
  How many children under the age of 18 do you have? _____

As noted earlier, when choosing the alternatives of response for each item, two conditions must be fulfilled. Categories should be (1) *exhaustive* (i.e., all possible responses must find a place in one of the options proposed); and (2) *mutually exclusive* (i.e., each response should correspond to only one precoded category, unless differently specified).

Pretesting the Questionnaire

Before administering a questionnaire to the actual sample of respondents, it is necessary to carry out at least one pretest (pilot test) to verify that it is well understood and does not yield obvious bias effects. The output of a pretesting phase can lead the researcher to (a) aggregate, specify, or better articulate the response alternatives, (b) revise or delete questions that raise many "I don't know," "I don't remember," "I don't want to answer" observations, specifications, explanations, or criticisms; (c) delete those questions that appear to have no variance; (d) integrate missing topics; (e) create a new order for the questions; and (f) verify the timing of the interview.

Pretests can be carried out in many ways: recording the reactions of the respondents during the interview, interviewing or debriefing the respondents *ex post*, asking the advice of a panel of experts, or mixing these methods. This phase should not be skipped, and a particular study could require more than one pretest before the questionnaire is finalized.

Administering the Questionnaire

Final questionnaires can be administered in many ways. In interviewer-administered questionnaires, particular care must be addressed to the training and monitoring of the interviewers. Another important distinction is the mode of data collection: face-to-face, telephone, mail, computer, and Internet data collection in their various forms (CATI, CAPI, CASI, etc.). Each one has its own advantages and drawbacks, in terms of costs, quickness, intrusiveness, anonymity guarantees, general design of the questionnaire, types of questions allowed and quality of the responses, return rates, and data entry. The researcher must find the right trade-off among costs and benefits with respect to all these factors.

Many new studies have focused on the role of the Internet in survey data collection. At least three big advantages characterize the use of the Internet in survey research: (1) the possibility to reach a huge population at a relatively low cost, (2) the possibility to minimize the intrusiveness of the interviewer and his or her instruments, and (3) the possibility to provide standardized multi-media stimuli (e.g., audio-visual content) as part of the questionnaire. These advantages to Web questionnaire notwithstanding, there also are many problems to take into account when using the Internet for data collection: the quality of the samples, verifying the eligibility of the respondents, the contexts in which questionnaires are completed, and so forth. Currently, it seems that a mix of traditional and innovative methods is necessary.

Finally, questionnaires can be administered on a single occasion (*cross-sectional design*) or at intervals over a period of time. In this second case, there are two alternatives: *trend* or *repeated cross-sectional* studies (when the same questionnaire is administered to different samples during the chosen period) and *longitudinal* studies (when the same questionnaire is administered to the same units). The most common type of longitudinal study is the *panel*, in which the same respondents are interviewed more than once, using a questionnaire that is identical, or at least very similar, to that used in each wave of the panel.

Alberto Trobia

See also Anonymity; Check All That Apply; Closed-Ended Question; Codebook; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Confidentiality; Construct Validity; Content Analysis; Contingency Question; Cover Letter; Cross-Sectional Survey Design; Double-Barreled Question; Double Negative; Exhaustive; Interval Measure; Interviewer Monitoring; Interviewer Training; Likert Scale; Longitudinal Studies; Mode of Data Collection; Mutually Exclusive; Nominal Measure; Open-Ended Question; Ordinal Measure; Panel Survey; Pilot Test; Precoded Question; Questionnaire Design; Questionnaire Length; Questionnaire-Related Error; Question Order Effects; Question Wording as Discourse Indicators; Ratio Measure; Reliability; Repeated Cross-Sectional Design; Respondent Debriefing; Respondent–Interviewer Rapport; Response Alternatives; Self-Administered Questionnaire; Semantic Differential Technique; Sensitive Topics; Validity; Wave

Further Readings

Alreck, P. L., & Settle, R. B. (1995). *The survey research handbook: Guidelines and strategies for conducting a survey.* Burr Ridge, IL: Irwin.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardised questionnaire.* London: Sage.

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464–494.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method.* New York: Wiley.

Foddy, W. H. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research.* Melbourne, Australia: Cambridge University Press.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Newell, R. (1993). Questionnaires. In N. Gilbert (Ed.), *Researching social life* (pp. 94–115). London: Sage.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., & Singer, E. (2004). *Methods for testing and evaluating survey questionnaires.* New York: Wiley.

Singer, E., & Presser, S. (1989). *Survey research methods: A reader.* Chicago: University of Chicago Press.

Sudman, S., & Bradburn, N. M. (1983). *Asking questions.* San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

QUESTIONNAIRE DESIGN

Questionnaire design is the process of designing the format and questions in the survey instrument that will be used to collect data about a particular phenomenon. In designing a questionnaire, all the various stages of survey design and implementation should be considered. These include the following nine elements: (1) determination of goals, objectives, and research questions; (2) definition of key concepts; (3) generation of hypotheses and proposed relationships; (4) choice of survey mode (mail, telephone, face-to-face, Internet); (5) question construction; (6) sampling; (7) questionnaire administration and data collection; (8) data summarization and analysis; (9) conclusions and communication of results.

One goal of the questionnaire design process is to reduce the total amount of measurement error in a questionnaire. Survey measurement error results from two sources: *variance* and *bias.* Question wording (including response alternatives), ordering, formatting, or all three, may introduce variance and bias into measurement, which affects the reliability and validity of the data and conclusions reached. Increased variance involves random increases in distance between the reported survey value from the "true" survey value (e.g., attitude, behavior, factual knowledge). Increased bias involves a systematic (directional) difference from the "true" survey value. Questionnaire-related error may be introduced by four different sources: the interviewer, the item wording/format/ordering, the respondent, and the mode of data collection. Careful questionnaire design can reduce the likelihood of error from each of these sources.

Goals, Objectives, and Hypotheses

The process of designing a questionnaire begins with a definition of the goals and objectives of the study. A clearly defined purpose acts as an anchor that sets the stage for questionnaire format and question construction or ordering. In what concepts and phenomena is the researcher interested? What does he or she want to learn? To what population are the results to generalize? These are all questions that should be asked during the initial phase of questionnaire design. Each item included in the questionnaire should meet the criteria that it will provide useful information related to the goals and research questions of the survey; parsimony is a virtue in questionnaire design. To properly define goals and research questions, key concepts of interest need to be detailed. A list of concepts of interest and how they relate to one another aids in selecting specific questions to include. These concepts are transformed into (i.e., operationalize as) survey questions that measure the concept in some way, proceeding from abstract concepts to specific measurements. Once the goals and objectives of the survey have been established, it is possible to generate specific research questions and, possibly, directional hypotheses based on theory and previous findings. The items used in the questionnaire should produce data that are appropriate to the desired analyses.

Questionnaire Format

The layout of a questionnaire, no matter what type, should reduce the cognitive burden of respondents (and interviewers) and contain an intuitive and logical flow. For example, in most cases, questions on related topics should be grouped together and questions should maintain the chronology of events. Questionnaire format should be as easy as possible to understand and use for the interviewer or the respondent depending on the mode of administration. Questions should be numbered individually, clearly spaced, and visually distinct from one another. It is especially important that self-administered questionnaires provide clear and concise instructions and have a simple layout (e.g., common question format, visual instructions for skipping questions). For interviewer-administered questionnaires, instructions that are to be read aloud to respondents should be visually distinct from instructions that are for only the interviewer; for example, set off by *italics*, CAPS, **bold type**, or parentheses ( ). Ideally, important questions should appear early in a questionnaire to avoid the possible negative effects of respondent fatigue on motivation, recall, and item nonresponse.

Questionnaires that have a professional appearance are taken more seriously by respondents (e.g., displaying professional affiliations). Social validation is also an important factor; questionnaires that end by thanking respondents for their time and effort and reminding respondents of the usefulness of the data they have provided are rated as more enjoyable by respondents.

Questionnaire length is an important issue for both cost reasons and effects on respondent behavior. First, longer questionnaires place a greater burden on respondents and are more costly. Second, longer questionnaires can have many significant effects on respondent behavior. Refusal rates rise with the length of the questionnaire; respondents may be willing to spend time on a 20-minute interview but may feel a 45-minute interview is far too long. In addition, greater attrition, measurement error, and satisficing may occur due to fatigue with longer questionnaires.

The contingencies that are built into a questionnaire, even when the questionnaire is computer assisted, must be designed with impeccable logic. Furthermore, when the questionnaire is not administered via computer and is self-administered (i.e., paper-and-pencil questionnaires), these so-called skip patterns must be explained in very clear and uncomplicated language, graphics, or both.

Question Wording

Unclear concepts, poorly worded questions, and difficult or unclear response choices may make the questionnaire difficult for both respondents and interviewers. Questionnaires should contain items that are both reliable and valid. Reliability is the consistency of the measurement; that is, the question is interpreted and responded to similarly over repeated trials. Construct validity is whether or not the measurement, as worded, properly reflects the underlying construct of interest.

Questions may be troublesome for respondents with respect to comprehension, retrieval, and response formation. First, the respondent must be able to understand the question. Questions should use simple (enough) terms and concepts that respondents are likely to know, and they should not contain vague and ambiguous terms; in other words, questions should be worded appropriately for the reading and knowledge level of the respondents. Questions should avoid unnecessary complexity, such as compound sentences, double-barreled questions (two questions within one), and double negatives. Longer questions that provide sufficient detail (without becoming convoluted) and are explicit generally will enhance comprehension.

Even though questions may be perfectly understandable for respondents, there may be problems with retrieval of the desired information from respondents' memory. Questions should ask for information that is as recent as possible and provide cues that match the original encoding context. Question construction may be either personal or impersonal in form: The choice depends on what kind of information, personal or general, the question is designed to elicit. Shorter recall periods (or reference periods) are easier for respondents to handle. Cues that are distinctive and personally relevant to the respondent also aid retrieval. Questions that ask for information that is general or was initially superficially encoded in the brain are prone to greater error in comparison to questions with appropriate, specific cues.

Terms that have multiple possible interpretations should have brief definitions provided in the question. Questions that define concepts for respondents work better if the definition of the concept precedes the actual question. Likewise, reference periods are also more effective when they are provided before the question. Event history calendars are useful for asking about similar events over a long period of time. It is also a good idea to avoid sensitive questions if possible (e.g., questions about income, religion, or sexual behavior). Any questions relating to sensitive information should be clearly necessary, and respondents should understand this need.

Response Alternatives

Responses can be either open-ended or closed-ended. *Open-ended* questions provide no predetermined response categories and allow the respondent to answer with whatever information he or she considers relevant. This answer format is more cognitively demanding for respondents, but it often results in more detailed and informative responses. Open-ended questions are good for gathering information on a topic for which there is no clear set of response categories. Open-ended questions work better with interviewer-administered questionnaires than in surveys that use self-administered questionnaires. An open-ended response format may result in a great deal of information, but it may be information that is not easily comparable or easily codable. Open-ended data also can be quite time-consuming and very costly to process.

Closed-ended questions ask respondents to select among a predetermined set of response categories. These response categories must be exhaustive and mutually exclusive. The closed-ended method reduces the cognitive burden of the respondent and enhances the ability to compare responses. The data are already coded (assigned a numerical value) and can be easily quantified, which saves data processing time and money. Closed-ended questions are ideal for self-administered questionnaires because they avoid the greater subjectivity and volatility of open-ended questions. However, if the researcher is not careful, the selection of response alternatives may bias respondents by framing thinking and by predetermining what are considered "appropriate" answers.

Response categories for event information may be divided by occurrence, absolute frequency (actual amount), relative frequency (quantifiers that denote more or less of something, such as "always," "most of the time," "sometimes," "basically never"), regularity, and specific dates. Attitude answers may take the form of ratings, rankings, agree–disagree scales, and forced choice. Attitude questions essentially require an evaluation of some phenomenon (e.g., approval/disapproval of the president). In selecting terms to use to describe attitudes, it is best to use terms that are simple and easy to understand, as well as terms that are likely to be interpreted similarly by all respondents.

Response scales require decisions concerning the length of scale, whether or not to include midpoint or explicit "don't know" options, and whether to use verbal labels, numbers, or both. A very common scale for closed-ended survey items is a Likert scale (e.g., "strongly agree," "somewhat agree," "neither agree nor disagree," "somewhat disagree," or "strongly disagree"). Rating scales can be presented as either a verbal description (e.g., strongly agree, strongly disagree), numbers (e.g., 1 to 5), or a combination of both. Meaningful verbal labels with relatively equal intervals provide a linguistic reference point for numerical scales, which significantly aids respondents.

Including a middle category is desirable when there is a readily identifiable and meaningful midpoint to a scale (e.g., increase, remain the same, decrease). The number of response options should attempt to maximize discrimination between response categories while maintaining the usefulness and meaningfulness of response categories. Research shows that five to nine response categories is best. There is tremendous debate over the usefulness of explicitly including "don't know" responses. Some argue that questionnaire designers are incorrect in assuming respondents always have an opinion and that surveys often force respondents to artificially produce an opinion. The lack of opinion regarding a topic has been referred to as displaying a "nonattitude." Others argue that explicit "don't know" response choices encourage respondents to satisfice and thereby not make the mental effort required to answer questions accurately. If given the chance, respondents with a low level of motivation will opt to get through a survey quickly and often choose the "don't know" response to facilitate this process, when otherwise they would have produced an opinion by not being offered a "don't know" option and thus been forced to put a little more effort into answering the question. A good situation in which to include an explicit "don't know" option is on unfamiliar topics where respondents are much more likely to have no knowledge or information about the subject and thus truly may have no opinion.

Question Testing Methods

Survey methodologists have developed several different methods of testing questionnaires to detect problems before they go out in the field. Question testing methods include the following: expert panels, traditional interviewer debriefing, behavior coding, cognitive interviewing, focus groups, and split-half experiments. Using one or some combination of methods improves the quality of a questionnaire.

Conventional pretesting involves interviewers conducting a relatively small number of interviews with a draft version of the questionnaire, followed up by an informal group debriefing session. Conventional pretesting accurately identifies problems that interviewers may have with the delivery of questions, but it is not as good at diagnosing respondent difficulties.

Behavioral coding standardizes the method of observing and recording the respondent–interviewer interaction. Behaviors that occur during the question–answer process are categorized (i.e., respondent asks for repetition of the response categories), and their frequency of occurrence is recorded. Behavioral coding aids in the identification of problems in question wording and question ordering and conceptual ambiguities. This method generates more reliability and consistency across interviewers in coding the question answering process because it uses a method that results in quantifiable data. One drawback to behavioral coding is that it does not yield any information for the *reasons* behind question problems.

Cognitive interviewing techniques are adept at revealing the largely invisible cognitive processes respondents engage in while answering questionnaires. Respondents may be asked to "think aloud" while answering questions, or respondents may be asked specific standardized question probes (e.g., "What did you count as 'exercise' when I asked you, *On average how many times do you exercise per week?*") designed to reveal the thinking process. This method is useful for identifying problems with question comprehension, recall strategies, sensitivity of questions, response selection, and vagueness of reference periods. The most common criticism of cognitive interviewing techniques is that they are somewhat artificial, sometimes causing the respondent to engage in behavior and thinking patterns they would not normally engage in when responding to a question.

Expert panels gather substantive and methodological experts to critique questionnaires. These panels are a relatively inexpensive and quick alternative to the other question testing methods. Experts have "seen it all" over the years and can easily point out troublesome aspects of questionnaires.

Focus groups, that is, small groups of respondents asked to discuss survey questions, are able to generate a large amount of information in a short amount of time. However, this technique sometimes may result in a skewed group consensus due to group dynamics.

Split-half experiments examine the effects of different question wording, ordering, or format by randomly assigning different forms of a question to different respondents. However, these tests can be costly and may require larger sample sizes than a survey otherwise needs.

Gregory G. Holyk

See also Attitude Measurement; Behavior Coding; Bias; Closed-Ended Question; Cognitive Interviewing; Construct Validity; Contingency Question; Event History Calendars; Exhaustive; Focus Group; Forced Choice; Interviewer Debriefing; Mutually Exclusive; Nonattitude; Open-Ended Question; Paper-and-Pencil Interviewing (PAPI); Questionnaire; Questionnaire Length; Questionnaire-Related Error; Question Wording as Discourse Indicators; Reliability; Respondent Burden; Response Alternatives; Satisficing; Validity; Variance

Further Readings

Aday, L. A. (2006). *Designing and conducting health surveys: A comprehensive guide* (3rd ed.). San Francisco: Jossey-Bass.

Dillman, D. A. (2006). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

Fowler, F. J., Jr. (1993). *Survey research methods* (2nd ed.). Newbury Park, CA: Sage.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Oppenheim, A. N. (1999). *Questionnaire design, interviewing, and attitude measurement.* New York: Pinter.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*, 65–88.

QUESTIONNAIRE LENGTH

Questionnaire length refers to the amount of time it takes a respondent to complete a questionnaire. Survey instruments can vary in length from less than a minute to more than an hour. The length of a questionnaire is important because it can directly affect response rates, survey costs, and data quality. Longer questionnaires result in higher data collection costs and greater respondent burden and may lead to lower response rates and diminished quality of response. However, practical experience and experimental data suggest that below a specific threshold, questionnaire length bears little relationship to response rate or data quality and has only a minor impact on survey costs.

For telephone and online surveys, although longer questionnaires will result in higher data collection costs, it is generally recognized that questionnaire length is a major factor for these modes only when it exceeds 20 minutes. The field cost of increasing the length of a telephone interview 1 minute beyond the 20-minute mark is 2 to 3 times as high as the cost of adding another minute to a 15-minute telephone interview. This higher cost reflects the lower response rates associated with longer interviews (due to the effects of the increase in respondent burden) and the greater number of interviewer hours needed to achieve the same sample size. The cost of administering a Web survey is similarly increased when the length of interview exceeds 20 minutes. In this case, additional monetary incentives and rewards must be offered to prompt respondents to complete a longer interview. For mail surveys, questionnaire length has a small impact on field costs compared to phone and online surveys. Longer mail questionnaires may require more postage and data entry will be more expensive, but these cost increases are generally not as large as those for other modes of data collection. With in-person interviewing, respondents do not appear to be as affected by the length of the questionnaire as they do by the length of other survey modes.

Questionnaire length also affects data quality. For telephone surveys, long questionnaires get lower response rates than short questionnaires and may therefore be subject to more potential nonresponse bias. The length of a telephone questionnaire affects response rates in two ways. First, when interview length is specified in the respondent introduction or otherwise disclosed by interviewers prior to the first substantive question, the number of refusals tends to be higher for a longer questionnaire than for a shorter one. Second, the more time an interview takes, the greater the chance that a respondent will break off the interview because of disinterest, to attend to personal business, or for some other reason. Refusals and break-offs can be recontacted in order to complete an interview, but this adds time and cost to a survey research project. There is also evidence that both respondents and interviewers can get fatigued toward the end of a long telephone survey, possibly compromising data quality, for example, through an increased tendency to satisfice. Mail survey data quality can also be affected by questionnaire length through lower response rates, but there is no widely accepted threshold beyond which a mail survey's response rate declines significantly.

As noted earlier, face-to-face interviewing is less affected by length of interview than are other methodologies, partly because respondents' initial agreement to allow interviewers into their homes tends to make them more receptive to making a major time commitment. Before telephone surveys became the most popular methodology, the norm was to conduct interviews face-to-face. The main cost incurred with in-person interviewing often was associated with sending the interviewer out into the field and not necessarily how long the interviewer remained at a particular household once he or she was there. These in-person surveys often ranged in interview length from 45 minutes to an hour. But even this methodology is subject to limits in questionnaire length; in-person surveys that take as long as 75 minutes to administer are reported to have lower response rates than those that can be completed in 45 minutes.

Questionnaire length is only one of many factors that influence a survey's data quality. For example, long questionnaires are less likely to have low response rates when respondents are engaged and interested in the survey topic, and long, mail survey questionnaires that are well-designed can result in higher response rates than shorter questionnaires that are not as attractive to respondents. By adding steps to the data collection process, such as advance letters, incentives, and reminder mailings or phone calls, researchers can match or even exceed the boost in response rate that would be obtained by shortening a questionnaire.

Larry Hugick and Jonathan Best

See also Partial Completion; Questionnaire; Questionnaire Design; Response Rates; Respondent Burden; Satisficing

Further Readings

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: Wiley.

Questionnaire-Related Error

Error related to the questionnaire is one component of total survey error. This type of error is traditionally viewed as being a discrepancy between the respondent's answer and the "true" score or answer. Questionnaire-related errors may stem from the content of the questionnaire (the wording, ordering, and/or formatting of the questions themselves), the method of delivery (whether by interviewer, by telephone, on paper, by computer, or some other mode), the materials that accompany the questionnaire (such as show cards in a face-to-face survey), or all of these things. The focus in this entry is primarily on response error by respondents, although these other factors can be considered secondary aspects of questionnaire-related error.

Recent work on questionnaire error draws on cognitive psychology to inform the design of standardized instruments. One prominent framework of the survey response process (or the mental steps a person undergoes when answering survey questions), advanced by Roger Tourangeau, Lance Rips, and Ken Rasinski, consists of four major components: (1) comprehension and interpretation of the question; (2) retrieval of relevant information from memory; (3) judgment, or deciding on an answer; and (4) response, including mapping the answer to response categories and reporting the answer.

Errors may arise at any step, and the questionnaire may affect the likelihood and extent of those errors. Although there are many design principles that minimize error, often there are trade-offs that researchers need to consider when adopting any given solution.

Comprehension

Taking steps to ensure question comprehension by respondents is often the first stage in developing a questionnaire that minimizes error due to misunderstanding or misinterpreting questions. In general, the more complex or ambiguous the task, the more error there will be. Questions may contain words unfamiliar to the respondent, may have complex syntax, or may assume a higher level of respondent cognitive ability than is warranted. Sometimes researchers use specialized terminology rather than colloquial language. Respondents may also not pay close attention and may miss part of the question.

One difficulty is that many words in natural language are ambiguous, and researchers often must use vague quantifiers that have an imprecise range of application. Some examples include categories of words denoting frequency (e.g., *never, sometimes, pretty often, very often*), probability expressions (e.g., *very unlikely, somewhat unlikely, likely*), amounts (e.g., *none, some, many*, *all*), and strength (e.g., *extremely dissatisfied, very dissatisfied, somewhat dissatisfied*). There is evidence that respondents have varied associations with words that denote frequency, regularity, or size. For example, *How often . . .?* results in variability based on how respondents interpret the word *often.* It is advisable to use more specific quantifiers, such as *On how many days . . .?* or *How many times in the last week/month/year . . .?*

Simple words or categories that appear to have common understandings, such as *household* or *sibling*, can be interpreted differently. When possible, such terms should be defined so as to minimize variability in how respondents interpret them. For example, *household* can include persons who use the address as their permanent address but who are not physically living there, such as college students who live in dormitories, individuals on military duty, or individuals who may reside at that address for only part of the year. *Sibling* may include biological siblings, half siblings, and adopted or stepsiblings.

Retrieval of Events and Behaviors

The retrieval stage of the response process involves bringing information from long-term memory to working memory. Though respondents occasionally may have the opportunity and the inclination to refer to documents for their answers, they most often need to rely on memory. In general, the more complex the memory task is, the less accurate the respondents' answers will be, thereby contributing to questionnaire-related error.

Most questions are time-bound so time frames or reference periods should be specified, partly to avoid ambiguity and partly to ease the recall task. For example, when asking how many people live in the household, the time span should be clear (e.g., *during the last 12 months, since [month] of last year*). Similarly, it should be clear if the reference period is calculated in days or is defined by the calendar (e.g., the last 30 days or the most recent calendar month). Longer reference periods increase both the time over which respondents have to remember events and the number of events respondents have to remember, both of which decrease accuracy. Because more recent events are easier to recall than older events, asking about more recent reference periods increases accuracy.

As the retention interval becomes longer, respondents rely less on direct recall and more on inference in answering questions. Thus respondents may be subject to *forward telescoping*, or mistakenly reporting events that happen before the reference period, as well as to *backward telescoping*, or omitting events that occur during the reference period. One method for reducing telescoping is *bounded recall*, which involves enumerating for respondents critical events that may have occurred at the boundaries for the reference period.

Questions asking for dates may be particularly difficult to answer accurately because dates (especially for common events such as going to the grocery store) are not normally kept in memory. Instead, respondents will infer dates from other life events they may use to anchor their memories (e.g., "in the spring, after my child was born . . ."). For example, college students may organize events in memory around the academic school year calendar (e.g., fall semester, homecoming weekend, Christmas break, spring break, freshman year). Questionnaire experts recommend using calendars and noting personal landmarks to promote recall. Distinctive or emotionally involving events (e.g., purchase of a new home) are easier to recall than nonsalient, frequent, and typical events (e.g., going to the ATM to withdraw money).

Smaller categories of events also help recall accuracy. For example, asking about time spent on housework during the past 7 days may be more manageable if respondents are asked separately about time spent cleaning, time spent cooking, and time spent washing and ironing (or other household activities). Providing multiple cues and decomposing the task can improve recall. In addition, spending more time on the task gives respondents more time to search memory. Thus longer introductions and a slower interviewing pace can be beneficial. Finally, some research suggests it is possible to motivate respondents to search their memories more carefully by highlighting the importance of accurate information; the impact of this, however, is normally small.

Threatening and Socially Desirable Behaviors

The likelihood and extent of respondent error may be correlated with the nature of the event being asked about and the respondent characteristics being measured. For example, a question asking respondents about the number of sexual partners they have had may be answered more accurately by respondents with few partners and less accurately by respondents with more partners because the latter have more to remember. In addition, this question may be sensitive; it may be perceived as socially desirable or undesirable to report a certain number of sexual partners. Questions about illegal behaviors, such as cocaine usage, may also be perceived as threatening and may result in underreporting of behaviors, whereas questions about behaviors perceived as socially desirable, such as voting, may result in overreporting of behaviors.

Minimizing error often involves reassuring the respondent of confidentiality, an action that encourages accurate reporting. However, while part of the error may be due to intentional misrepresentation, the social desirability of a behavior may interact with memory so that socially undesirable behaviors are remembered as occurring less frequently, and socially desirable behaviors are remembered as occurring more frequently. Thus memory aids can increase accuracy in the reporting of such behaviors.

Subjective Phenomena Questions

As in questions about events and behaviors, accuracy for subjective questions necessitates comprehension on the part of the respondent. Subjective questions may be poorly worded or confusing, leading to variability in respondent interpretations of certain words or phrases. Questions may also be double-barreled, making it difficult to respond to them. However, while events and behaviors generally are thought to entail recall from long-term memory, questions asking about attitudes involve a combination of memory and inference.

Traditionally, attitudes were thought to be long-term values that were recalled from memory at the moment when asked. Recent research on attitudes suggests that there may be instances when attitudes are recalled from memory and other instances when they are formed anew each time they are called for. That is, rather than search the mental "file cabinet" for one's rating of the president or one's attitude on foreign policy, respondents integrate the bits of information in working memory into an attitude, for example, each time they are asked about the president or foreign policy. Respondents may use previous evaluations, impressions, general values, specific beliefs, or even information from prior survey questions to form their responses.

Thus the notion of accuracy in reports of subjective phenomena differs from accuracy in reports of events and behaviors. Accuracy often means that responses were not unduly swayed in one direction, that the question and response alternatives were balanced, and that biasing information was not conveyed to the respondent. In addition, the context in which the question was asked (e.g., what other question or questions closely preceded it) is important to how respondents answer because information provided earlier may frame the context for current questions.

When researchers report the results of attitude questions, they must make the question wording and context available to the reader. It is impossible to talk about accuracy or error without knowing the wording and context in which the question was asked. Furthermore, results related to attitude stability or reliability need to be interpreted with caution. Such statistics may help point out questions that need to be improved, but low stability or reliability scores are not necessarily artifacts of measurement. They may also be substantively informative about people's attitudes on a given topic.

Trade-Offs

Although the questionnaire design principles mentioned earlier minimize many errors, they may introduce yet others and need to be balanced against other considerations, namely, interview time, respondent burden, and cost. The more strategies that are used to help respondents remember, the more likely it is that the questionnaire will encourage recall of events and behaviors that occur outside the time frame of interest. Similarly, providing examples may help the respondent understand what is meant by the question, but it may focus the respondent exclusively on the examples provided. In asking how many days the respondent reads a news magazine, one might provide examples (e.g., . . . *such as* Time, Newsweek, *or* US News and World Report) at the risk of focusing the respondent entirely on the examples and excluding other news magazines that are not in the question text. Thus specificity in the question may reduce error due to comprehension or interpretation problems, but it may introduce other types of error, such as errors of omission.

In addition, many of the solutions to increase accuracy also increase cost. Providing examples in the question text lengthens the question; decomposing all events and behaviors in a questionnaire could lengthen the interview considerably; and using event history calendars adds extra time to administer. When multiplied over many questions, this may increase questionnaire length and respondent burden. When multiplied over many respondents, this can increase costs. Researchers must decide which trade-offs are most appropriate for their particular survey. In addition, the more tasks there are for interviewers to administer, the more interviewer variability is introduced as a source of error.

Questionnaire design that minimizes error is often thought of as being part art and part science. Good questionnaires generally combine good design principles (the "art") and insights from cognitive psychology (the "science"). Researchers continue to strive to make questionnaire design more of a science; however, there are many particularities in how people remember and report events and subjective experiences in different areas of their lives, and there is unfortunately no overarching theory of questionnaire design. In general, it is good to invest resources in developing appropriate questions using ethnographic techniques, in evaluating comprehension and interpretation using cognitive interviewing, and in testing and evaluating questionnaires prior to fielding them. Pilot surveys are invaluable and provide an opportunity for behavior coding to systematically count the occurrence of problematic interviewer or respondent behaviors. Finally, interviewing protocols and other aspects of instrumentation (such as visual layout of a questionnaire) also need to be considered.

Danna Basson

See also Bounding; Cognitive Aspects of Survey Methodology (CASM); Cognitive Interviewing; Confidentiality; Context Effect; Double-Barreled Question; Errors of Omission; Interviewer-Related Error; Measurement Error; Mode-Related Error; Pilot Test; Questionnaire Design; Questionnaire Length; Question Order Effects; Reference Period; Respondent Burden; Respondent-Related Error; Social Desirability; Telescoping; Total Survey Error (TSE); True Value; Visual Communication

Further Readings

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design for market research, politics, and social and health questions.* San Francisco: Jossey-Bass.

Presser, S., Rothgeb, J. M., Couper, M. C., Lessler, J. T., Martin, E., Martin, J., et al. (2004). *Methods for testing and evaluating survey questionnaires.* Hoboken, NJ: Wiley-Interscience.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments in questions form, wording, and context.* Thousand Oaks, CA: Sage.

Sirken, M. G., Hermann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Tourangeau, R. (Eds.). (1999). *Cognition and survey research.* New York: Wiley.

Tourangeau R., Rips, L. J., & Rasinski, K. (2000). *The psychology of the survey response.* Cambridge, UK: Cambridge University Press.

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions or revealing preferences? *American Journal of Political Science*, *36*, 579–616.

Question Order Effects

The order in which questions are asked in a questionnaire can have a significant effect on the results. The preceding questions provide the context in which the respondent answers an item, and changing this context can make a large difference in the survey results.

There are a number of ways in which items that appear earlier in a questionnaire can affect responses to later questions. One is by establishing a norm of reciprocity or fairness, a frequently cited example of which is provided by the work of Herbert Hyman and Paul Sheatsley from more than 50 years ago. These researchers varied the order of two questions: one on whether the United States should allow communist reporters from other countries to come to the United States and send back to their newspapers the news as they saw it and another on whether a communist country like Russia should let American newspaper reporters come in and send back to America the news as they saw it. Changing the order resulted in a difference of 37 percentage points in the percentage of "yes" responses to the question on communist reporters and a 24 percentage point difference in the results for the American reporters item. When either of the items was asked second, the context for the item was changed as a result of the answer to the first, and the responses to the second were more in line with what would be considered "fair," based on the previous response.

Another way in which the earlier questions in a survey can affect results is by altering the frame in which a question is interpreted. For example, if respondents are asked about their interest in politics as the first item in a series, their reported level of interest is likely to be higher than if this question appears after a series of potentially difficult political knowledge items, such as whether they happen to remember anything special that their U.S. Representative had done for their district or how they had voted on any particular legislative bill. When the "interest in politics" item is asked first, respondents answer in terms of their general experience and may report a fairly high level of interest based on watching the network news regularly or talking about politics with their friends. Asking the knowledge questions first changes the meaning of "interest in politics"; respondents are more likely to interpret this in terms of having specific knowledge about how their member of Congress has acted. Because most respondents cannot remember any such specific information, the context of the question leads them to report a lower level of political interest.

Question order can also change the salience of various alternatives. In surveys that ask about the most important problem facing the country, for example, this question generally will be one of the first items asked. In this position, respondents are more likely to provide an answer based on their recent experience. If they are first asked a series of questions about some issue such as immigration or public education, the percentage of respondents who cite this issue as "the most important problem" will be higher. By making immigration or public education more salient to the respondent, this question order will change the context in which they interpret the "most important problem" item, and these different question orders will produce different results.

Another type of question order effect that has been identified is termed a *part–whole contrast effect*. Such an effect can occur when a series of items on a particular topic includes both a rather general item and a more specific question. When the general and specific items are asked in different orders, the results for the specific item are generally unaffected, whereas those for the general item change significantly. In these situations, agreement with the more general item implies agreement with the more specific item, but the reverse is not true. When respondents are asked the general item after the more specific one, they may "subtract" their response to the specific question, altering the distribution of responses to the general item when it is asked after, rather than before, the more specific one.

Series of questions that have the same response format are also subject to effects due to question order. Items in such series may be affected by the content of the items that precede them in the list or potential response set bias among respondents. For example, in a survey about city services in which respondents are asked to rate various services as either excellent, good, fair, poor, or very poor, ratings for garbage collection services may differ depending on whether this item is asked before or after other services such as police protection, fire protection, or arts and cultural activities. One way in which researchers have attempted to address this issue is through randomization of these items, in which each of them is presented first to some respondents, each presented last to others, and at all positions in between across respondents. Through randomization, all possible orders of this set of items may be presented. Although this technique does not necessarily eliminate potential question order effects, randomization is thought to distribute any such effects across the set of items, so that no particular item is either relatively advantaged or disadvantaged by its position in the list.

There are no technical procedures for eliminating question order effects. Not every question can be first in the questionnaire, and questions asked early set the context in which later questions are interpreted. Question order effects are also persistent, and separating items for which there is likely to be an order effect with questions on unrelated topics generally does not eliminate such effects. In designing any questionnaire, therefore, careful consideration must be given to the potential for question order effects. Conducting cognitive interviews and a thorough pretest of the questionnaire also will help to identify such possible effects. Procedures such as randomization of sets of items or systematically varying the order of questions will serve to minimize the impact of question order or enable the researcher to specify the magnitude of such effects.

It is difficult to identify in advance the contexts in which question order effects will occur, and some effects will certainly arise despite the best efforts of the designers of survey questionnaires to minimize them. Identifying such effects is important in any analysis, but it may be of particular consequence in trend analyses, in which the focus is on change in opinions or attitudes. Even when the identical question is asked at different points in time, varying results may be due not to actual change but rather to the different order (context) in which the questions were asked.

Robert W. Oldendick

See also Cognitive Interviewing; Context Effect; Pilot Test; Random Order; Random Start; Response Bias; Topic Saliency

Further Readings

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1984). What must my interest in politics be if I just told you "I don't know"? *Public Opinion Quarterly*, *2*, 510–519.

Hyman, H., & Sheatsley, P. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The teaching of public affairs: Twenty-first Yearbook of the National Council of Social Studies* (pp. 11–34). Washington, DC: National Council for the Social Studies.

Schuman, H., Kalton, G., & Ludwig, J. (1983). Context and contiguity in survey questionnaires. *Public Opinion Quarterly*, *1*, 112–115.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys.* Thousand Oaks, CA: Sage.

Sigelman, L. (1981). Question-order effects on presidential popularity. *Public Opinion Quarterly*, *2*, 199–207.

Question Stem

A question stem is the part of the survey question that presents the issue about which the question is asking. With closed-ended questions, the stem can perhaps best be defined as the first half of a survey question that consists of two parts: (1) the wording or text that presents the issue the respondent is being asked to consider (along with any instructions, definitions, etc.) and (2) the answer options (response alternatives) from which a respondent may choose.

Survey researchers must strive to craft question stems that meet a number of important criteria. First and foremost, question stems must be written so that, to the degree that this can be controlled, all respondents understand the question being posed to them as meaning the same thing. If a given question is perceived differently by different types of respondents, it essentially becomes multiple questions capturing different things, depending on respondent perception. Because of this, survey researchers must make every attempt to word question stems simply and write to the population of interest. Question designers should consider providing definitions of terms and any guidance they can in a given question stem to ensure uniform understanding of concepts and terms.

Although questions including definitions, examples, and so forth, are created with the best of intentions, they may be too cumbersome for respondents. Survey researchers should avoid question stems that are too long and wordy; question stems should be as succinct as possible. Survey researchers are encouraged to break up complex issues into separate questions in order to aid respondent understanding. In doing so, the creation of double-barreled questions is also avoided.

In addition to being worded to ensure respondents' understanding, question stems must also be carefully worded to avoid affecting the results of a given item. Question stems should be written without the use of inflammatory phrases and should not purposely contain emotional or slanted language that could influence respondents and bias results. As the survey research industry continues to move toward the use of mixed-mode data collection and the full conversion of surveys from one mode to another, question stems must be checked carefully for mode appropriateness as well. Different modes rely on different words, instructions, and phrasing to communicate ideas, and question stems must be constructed in ways that do not confuse respondents. For example, closed-ended question stems offered in mail surveys, where respondents are able to see answer choices below question text, often must be altered for telephone surveys, where the categories must become part of the question stem that is read to respondents.

Although it is simple to define a question stem in theory, survey researchers may find it difficult in practice to develop a question stem that is appropriate, adequate, and understood by respondents. Many attempt to develop high-quality question stems in an effort to capture information on a respondent's attitude or beliefs toward a given issue but fall short due to a number of pitfalls. For this reason, experts have encouraged researchers for decades to design question stems based on quality items already developed and tested.

David James Roe

See also Bias; Closed-Ended Question; Mode of Data Collection; Open-Ended Question; Questionnaire Design

Further Readings

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording and context.* Thousand Oaks, CA: Sage.

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Question Wording as Discourse Indicators

Many of the questions used by survey researchers serve to solicit people's opinions on topical issues, and the distribution of the answers to survey questions makes up what is usually considered to be the "public's opinions." Thus, most research on survey questions treats these questions as "stimuli" and focuses on their communicative and evocative functions. As such, scholars examine response biases associated with certain question formats or wordings, the cognitive processes underlying these biases, and the flow and conversational logic of the interviewing process. But survey questions also may be viewed as responses, where those who formulate the questions are responding to meaningful social forces and conditions. Thus, survey questions can be excellent indicators for public discourse.

From this perspective, *the wording of survey questions* becomes the focus of analysis, and when studied systematically over time, question wording indicates the evolving discourse packages of important public issues. This entry introduces the notion of question wording as discourse indicators, develops its rationale, proposes content-analytic methods and concepts for such discourse analysis, and provides empirical examples of this approach.

Survey Questions as Indicators of Social Discourse

Public opinion and public discourse are often viewed as two parallel systems interacting with each other to provide interpretation and meaning to relevant events and policies. Discourse analysis reveals the evolving culture of issues and events in terms of their interpretive frames. The analysis of public opinion provides a glimpse into how individuals adopt such frames and which perspectives prevail on the aggregate. Frequency distributions of public opinion polls provide the latter. Public discourse packages of important public issues can be captured extremely well by content analysis of questions in public opinion polls through systematic examination of the language of survey questions.

It is commonly accepted that question–answer situations must be understood in the social context in which they take place. This contextual approach applies to the interview setting as well as to the content and form of the questions themselves, which must be seen as segments of the ongoing social discourse. In fact, survey questions and the way they frame issues are increasingly viewed as part of the elite discourse and political debate, which effectively shape popular discourse and thinking. Pollsters who design surveys are in the same position as public officials, editors, journalists, and newsmakers of all sorts: All choose how public issues are to be framed, and their choices have consequences.

Indeed there are good reasons to think of survey questions as useful indicators of social discourse on a specific issue. First, to be valid and reliable and to engage respondents, thus not adding to nonresponse, survey questions must be formulated in a way that carries widely shared meaning. Thus, pollsters often use words that respondents would use themselves in the same context, or they define key concepts in the questions in terms familiar to the respondents. Either way, the choice of wording reflects the ongoing political debate and, eventually, public discourse. Second, public opinion surveys and polls tend to focus on topical issues and current events and are likely to share similar choice criteria used by editors to decide what is news. Just like news items, they tend to concentrate on salient events and to reflect critical discourse moments. Third, when pollsters change question wording—often a hard decision for pollsters to make—they usually intend to either adapt to meaningful social changes or to write a better question. But writing better questions often means simply fine-tuning the questions so as to fit into a frame of reference perceived, by the pollster, as dominant or most relevant. Sometimes changes in questions are made to increase respondents' capability or willingness to respond. These changes do not occur arbitrarily but often indicate pollsters' assumptions about the salience of an issue, the public's level of knowledge, and the prevailing social norms affecting willingness to respond. In short, the reasons for introducing changes in question form and wording may, in fact, reflect developments in public discourse.

In addition to changes in question wording over time, there also may be differences in the wording and format used by separate research organizations and distinct studies of the same issue at a given point in time. These differences are often seen as nuisances by researchers who attempt to grasp public opinion on an issue. But for discourse analysis, these variations only enrich knowledge of the competing discourse frames and packages they may represent. Moreover, when such differences diminish and pollsters converge on similarly worded questions, this may indicate the emergence of a dominant discourse frame.

Content Analysis of Survey Questions

Various content-analytic methods and concepts may be applied to discourse analysis of question wording. Linguistic approaches offer especially profound and valuable tools for grasping the meaning embedded within the text without compromising reliability and generalizability. One such approach is that of Roberto Franzosi, which utilizes a semantic text grammar to organize textual data around a simple but flexible structure of Subject-Action-Object and their modifiers. The scheme reliably organizes textual data both hierarchically and relationally, with explicit links between important elements of discourse such as actors and actions. It is most useful for analyzing changes in the role and type of actors over time, in the nature of actions, and in their relationships with other elements of text. This fits the structure of many survey questions, which present respondents with alternative situations, scenarios, positions, or policies involving actors and a range of possible actions.

In addition to *semantic text grammar coding*, syntactic features that define the surface structure of the text (e.g., the use of passive or active voice) may provide valuable discourse indicators for the ideological leaning and cultural biases embedded in texts.

The deep structural schema of survey questions differs from that of other texts such as news reports. Importantly, even more than other types of text, standardized survey questions must presuppose respondents' social knowledge and integrate or supplement that knowledge in their wording. Coding question presuppositions offers especially valuable and compact indications of the "taken for granted" in public discourse, mutually known to all participants and often linked to ideological stances.

An illustration of the use of survey questions as discourse indicators may be given from a study of Israeli discourse on the Israeli–Arab conflict. Jacob Shamir, Neta Ziskind, and Shoshana Blum-Kulka built a comprehensive database of questions on peace and the territories. The data were collected systematically from the principal polling organizations, academic studies, and the leading daily newspapers. The study covered a 25-year period from June 1967, immediately after the Six-Day War, to October 1991 when the Madrid peace conference was convened. The study found a self-centered and self-serving discourse, which viewed Israel as the key actor in the conflict. Arab states and Palestinians were mainly action recipients and sometimes secondary initiators. Most noticeable was the marginal role of the Palestinian actor in Israeli discourse. Over time, the analysis revealed two major shifts in discourse on the territories. One was a transition from a general, abstract, and simplistic "territories for peace" frame to a concrete and policy-oriented approach referring to specific policy options. At the same time, the focus of discourse frames shifted from land to people: from a virtually exclusive focus on the territories to greater awareness of the Palestinian inhabitants in the territories. These two changes were not unrelated and coincided with a gradual increase in attention in the discourse to conflicting values in the political culture.

Summary

Public opinion surveys provide a natural arena in which discourse frames thrive and can be traced. Surveys closely follow current events and issues in the news. They constitute a key discourse domain of ongoing social action, a kind of self-contained system of communication highly indicative of the world of political discourse in general. They provide critical historical instances in the form of mini-texts, in which major themes on the public agenda, assumptions, and ideologies are crystallized in the form of questions. Therefore, content analysis of question wording provides an appealing approach for the study of political discourse.

Jacob Shamir

See also Content Analysis; Poll; Pollster; Public Opinion; Questionnaire; Questionnaire Design

Further Readings

Franzosi, R. (2004). *From words to numbers: Narrative, data, and social science.* Cambridge, UK: Cambridge University Press.

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, *9*, 1–37.

Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate with survey questions: The case of white opinion on affirmative action for blacks. *Social Cognition*, *8*, 73–103.

Shamir, J., Ziskind, N., & Blum-Kulka, S. (1999). What's in a question: A content analysis of survey questions. *Communication Review*, *3*, 353–377.

QUOTA SAMPLING

Quota sampling falls under the category of nonprobability sampling. Sampling involves the selection of a portion of the population being studied. In probability sampling each element in the population has a known nonzero chance of being selected through the use of a random selection procedure such as simple random sampling. Nonprobability sampling does not involve known nonzero probabilities of selection. Rather, subjective methods are used to decide which elements should be included in the sample. In nonprobability sampling the population may not be well defined. Nonprobability sampling is often divided into three categories: purposive sampling, convenience sampling, and quota sampling.

Quota sampling has some similarities to stratified sampling. The basic idea of quota sampling is to set a target number of completed interviews with specific subgroups of the population of interest. The sampling procedure then proceeds using a nonrandom selection mechanism until the desired number of completed interviews is obtained for each subgroup. A common example is to set 50% of the interviews with males and 50% with females in a random-digit dialing telephone interview survey. A sample of telephone numbers is released to the interviewers for calling. At the start of the survey, one adult is randomly selected from a sample household. It is generally more difficult to obtain interviews with males. So for example, if the total desired number of interviews is 1,000 (500 males and 500 females), and interviews with 500 females are obtained before interviews with 500 males, then no further interviews would be conducted with females and only males would be randomly selected and interviewed until the target of 500 males is reached. Females in those sample households would have a zero probability of selection. Also, because the 500 female interviews were most likely obtained at earlier call attempts, before the sample telephone numbers were thoroughly worked by the interviewers, females living in harder-to-reach households are less likely to be included in the sample of 500 females.

Quotas are often based on more than one characteristic. For example, a quota sample might have interviewer-assigned quotas for age by gender by employment status categories. For a given sample household the interviewer might ask for the rarest group first, and if a member of that group is present in the household, that individual will be interviewed. If a member of the rarest group is not present in the household, then an individual in one of the other rare groups will be selected. Once the quotas for the rare groups are filled, the interviewer will shift to filling the quotas for the more common groups.

The most famous example of the limitations of this type of quota sampling approach is the failure of the pre-election polls to predict the results of the 1948 U.S. presidential election. The field interviewers were given quotas to fill based on characteristics such as age, gender, race, degree of urbanicity, and socioeconomic status. The interviewers were then free to fill the quotas without any probability sampling mechanism in place. This subjective selection method resulted in a tendency for Republicans being more likely to be interviewed within the quota groups than Democrats. This resulted in the sample containing too many Republicans and causing the pre-election polls to incorrectly predict Thomas Dewey (the Republican candidate) as the winner.

Quota sampling is sometimes used in conjunction with area probability sampling of households. Area probability sampling techniques are used to select primary sampling units and segments. For each sample segment (e.g., city block) the interviewer is instructed to start at a corner of the segment and proceed around the segment, contacting housing units until a specific number of interviews are completed in the segment.

A major problem with quota sampling is the introduction of unknown sampling biases into the survey estimates. In the case of the 1948 U.S. presidential election, the sampling bias was associated with too many Republicans being selected. Another problem with quota sampling is that the sampling procedure often results in a lower response rate than would be achieved in a probability sample. Most quota samples stop attempting to complete interviews with active sample households once the quotas have been met. If a large amount of sample is active at the time the quotas are closed, then the response rate will be very low.

Michael P. Battaglia

See also Convenience Sampling; Nonprobability Sampling; Probability Sample; Purposive Sample; Simple Random Sample; Stratified Sampling

Further Readings

Henry, G. (1990). *Practical sampling.* Newbury Park, CA: Sage.

Kish, L. (1965). *Survey sampling* (2nd ed.). New York: Wiley.

R

Radio Buttons

A radio button is a type of survey response format used in electronic questionnaire media such as Web surveys, email surveys, personal digital assistant (PDA) applications, and other electronic documents. The radio button response format allows respondents to select one, and only one, of a set of two or more mutually exclusive response options. Respondents select the response option of interest by using a mouse pointer, keyboard key, or touch-screen stylus. The term *radio button* is a reference to the punch-in buttons invented years ago on radios to choose a preset station, such as in an automobile; as each new button is pressed, the previously pressed button is returned to the neutral position.

Radio buttons are often displayed as a series of open circles with a value or value label shown next to each button. They can be listed vertically, horizontally, or in a grid or matrix. Other electronic response formats include (a) drop-down boxes for longer lists of options, (b) check boxes for check-all-that-apply questions, and (c) text input fields for open-ended responses. Radio buttons, like other on-screen electronic response formats, require the respondent to exercise care when selecting the option of interest, making the format susceptible to careless respondent error. This is particularly true for respondents with less extensive computer experience.

Although drop-down boxes require less screen space than do radio buttons, the latter more closely resemble the format used in paper-based questionnaires. Additionally, while all radio button options are immediately visible to the respondent, drop-down boxes require a series of actions by the respondent to open the drop-down list, scroll through the full set of options, and select a response. This difference can result in greater respondent burden, increased cognitive difficulty for the respondent, and response distribution differences due to primacy effects.

Adam Safir

See also Drop-Down Menus; Internet Surveys; Mutually Exclusive; Primacy Effects; Questionnaire Design; Respondent Error; Web Survey

Further Readings

Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in Web surveys. *Social Science Computer Review*, *20* (4), 471–484.

Tourangeau, R., Crawford, S., Conrad, F. G., & Couper, M. P. (2004). What they see is what we get: Response options for Web surveys. *Social Science Computer Review*, *22* (1), 111–127.

Raking

Raking (also called raking ratio estimation) is a post-stratification procedure for adjusting the sample weights in a survey so that the adjusted weights add up to known population totals for the post-stratified classifications when only the marginal population totals are known. The resulting adjusted sample weights provide a closer match between the sample and the population across these post-strata than the original sample. Raking, however, assumes that nonrespondents in each post-stratum are like respondents. Nonetheless, when implemented with care, raking improves the mean squared error of sample estimates.

The term *raking* is used to describe this statistical technique because the raking ratio—the ratio of the population total (or control) total for a given post-stratum to the marginal row (or column) total from the sample for that same post-stratum—is calculated and then applied to each of the cells in that row (or column). This is done for each of the post-strata and repeated iteratively multiple times until the marginal row and column totals converge to the population totals. In essence, the raking ratio is "raked" over the cells in the respective rows and columns until convergence to the population totals is achieved, hence the term.

Raking was developed by W. Edwards Deming and Frederick F. Stephan in the late 1930s and used with the 1940 U.S. Census to ensure consistency between the census and samples taken from it. The computational procedure is the same as *iterative proportional fitting* used in the analysis of contingency tables, but the latter is not typically formulated in terms of weight adjustment.

The Two-Variable Raking Procedure

In two-variable (or two-dimensional) raking, population totals are known for the strata of two distinct variables. This situation can be represented by a rectangular array of cell estimates based on the initial sample weights with the "true" population row and column totals, called the row and column control totals, known. One objective is to adjust the cell entries so that both the row and column sums of the cell entries add up to the control totals.

The initial sample weights reflect the probabilities with which the sample units were selected and may incorporate a nonresponse adjustment as well. To preserve the original sample design and possible nonresponse adjustment, one objective of the raking procedure is to change the initial cell estimates as little as possible, subject to their adding up to the control totals in both dimensions.

The steps in raking are as follows. For the first variable (the row variable in a two-dimensional cross-tabulation), multiply the first row by the control total for row 1 divided by the sum of the cell entries for row 1. This adjustment will now make the cell entries for row 1 sum to the control total for that row. Then do the corresponding adjustment (multiplying the appropriate raking ratio for row 2 by each row 2 entry) to row 2. Continue in this way until all rows have been adjusted with each row summing to its respective control total.

For the second variable, multiply the first column by the control total for column 1 divided by the sum of the cell entries for column 1. At this point column 1 adds up to the control total for column 1, but the rows may no longer add up to their respective control totals. Continue multiplying each column by its respective raking ratio until all columns add up to their respective control totals. Then repeat the raking ratio adjustments on the rows, then the columns in an iterative fashion. (The appropriate raking ratios for the rows and columns will likely change in each iteration until the process converges.) Raking stops when all the rows *and* columns are within a pre-specified degree of tolerance or if the process is not converging. If raking converges, the new sampling weight for a sampling unit in a particular cell is the initial sampling weight times the raking-adjusted cell estimate divided by the initial cell estimate.

Because raking adjusts by a series of multiplications, any zero cell entry will remain zero. This property is advantageous in the case of structural zeros: cells that must be zero by definition or that are known to be zero in the entire population. In other cases, if the initial sample estimate for a cell is zero, but the population value may be positive, for example, the number of people suffering from a rare disease in a geographic area, it may be advisable to replace the initial zero estimate for the cell by a small positive value.

For two-variable raking, the process is known to converge if all initial cell entries are nonzero. If there are zero cells, for some configurations the process will not converge.

The Many-Variable Raking Procedure

Raking can also be performed when there are more than two distinct variables. For three variables, one has control totals for rows, columns, and pillars. The rows are adjusted to add up to the control totals, then the columns, then the pillars. These raking steps are repeated until convergence occurs. Raking is similarly specified for more than three control variables. Convergence is difficult to ascertain in advance, but the presence of zero cells is an important factor.

The Closeness Property

One of the properties of raking is that the raking-adjusted cell estimates are as close, in a specific sense, to the initial cell estimates as is possible subject to the raking-adjusted cell estimates adding up to the controls. For two-variable raking, let c_{ij} and r_{ij} denote, respectively, the initial and raking-adjusted cell estimates for the cell in row i and column j. Early researchers speculated that raking minimizes (subject to the constraints of the control totals) the weighted least squares distance between the c_{ij} and r_{ij}. In fact, raking minimizes (subject to the constraints of the control totals) $\sum_i \sum_j r_{ij} \log(r_{ij}/c_{ij})$ where the sum is over all cells ij with $c_{ij} > 0$. It can be verified that the sum is always nonnegative and only zero when $r_{ij} = c_{ij}$ for $c_{ij} > 0$. The corresponding conditions hold for more-than-two variable raking.

Practical Considerations

Raking is often employed when the control totals are not known exactly but can be estimated with higher accuracy than can the cell estimates. For example, for household surveys in the United States, demographic estimates from the Current Population Survey or the American Community Survey may be deemed accurate enough to use as control totals.

Much experience and fine tuning are needed in raking to determine the number of control variables to use and the number of cells. Too many control variables may lead to convergence problems or highly variable weights resulting in increased variance. In addition, if the variables used in the raking are not highly correlated with all the variables in the sample, then the variances for the uncorrelated variables can increase. Too few control variables may result in sample estimates that do not match the population as closely as otherwise could be achieved.

Michael P. Cohen

See also Bias; Mean Square Error; Post-Stratification; Weighting

Further Readings

Bishop, Y. M. M., Feinberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge: MIT Press.

Deming, W. E. (1943). *Statistical adjustment of data*. New York: Wiley.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 427–444.

Kalton, G. (1983). *Compensating for missing survey data*. Ann Arbor: University of Michigan, Institute for Social Research.

Random

A process is *random* if its outcome is one of several possible outcomes and is unknown prior to the execution of the process; that is, it results merely by chance. Some survey operations are carried out as random processes because of advantages resulting from multiple possible outcomes or from the absence of advance knowledge about the outcome. Such survey operations are said to be *randomized* and are discussed in more detail later in this entry. *Random numbers* are the realized outcomes of a numeric-valued random process. They are used by researchers or survey-operations staff to carry out randomized survey operations. This entry discusses properties and sources of random numbers.

For some randomization operations, the needed random numbers must be integers uniformly distributed between 1 and some maximum value. For other operations the needed random numbers must be uniformly distributed between 0 and 1. Statistical tests exist for testing the quality of random numbers. These include tests for goodness of fit to the uniform distribution, tests for the absence of trends in sequences of random numbers, and tests for the absence of nonzero correlations between pairs of random numbers at fixed lags. Random numbers obtained by using simple devices such as dice, slips of paper drawn from a hat, or numbered balls removed from an urn often fail one or more of these tests and should not be used in professional survey work.

Tables of random numbers appearing in textbooks about statistical methods or survey sampling procedures can be used as sources of small quantities of random numbers. Early tables of random numbers were created by using series of digits from an irrational number, such as e or π, or by using a special-purpose electronic “roulette wheel.” A much easier way to create tables of random numbers and to produce large quantities of random numbers for automated operations is to have a computer program generate *pseudorandom numbers*. These are numbers that appear to be random but are actually deterministic. A computer program cannot generate true random numbers because a program's outputs are completely determined by its logic and inputs, so outputs from a computer program are known. Algorithms for generating pseudorandom numbers appear easy to program, but there are a number of machine-dependent considerations and choices of parameters associated with successful implementation, so it is usually better to use existing computer code that has been thoroughly evaluated.

Richard Sigman

See also Random Assignment; Random Order; Random Sampling; Random Start

Further Readings

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods* (2nd ed.). New York: Springer.

Random Assignment

Random assignment is a term that is associated with true experiments (called controlled clinical trials in medical research) in which the effects of two or more "treatments" are compared with one another. Participants (respondents, subjects, etc.) are allocated to treatment conditions in such a way that each participant has the same chance of being a member of a particular treatment group. The groups so constituted are therefore as similar as possible at the beginning of the experiment before the treatments are introduced. If they differ at the end of the experiment, there will be a statistical basis for determining whether or not, or to what extent, it is the treatments that were the cause of the difference.

Traditionally, survey research and experimental research have not had much in common with one another. However, the following examples illustrate some of the many ways that survey researchers can benefit from random assignment and experimental design.

A survey researcher would like to determine whether it would be better to use answer sheets for a particular questionnaire in which the responses to individual items are given by blackening in small ovals (the "new" way) or by blackening in the spaces between pairs of vertical dotted lines (the "old" way). In a pilot study carried out prior to the main study, two versions of the answer sheets would be prepared and participants would be assigned to one version or the other in such a way that chance, and chance alone, determines which form they receive. This illustrates *simple* random assignment. (The effect of type of answer sheet might be operationalized by the difference between the percentages of omitted items for the two forms, since large numbers of omitted items are not desirable in survey research.)

For *stratified* random assignment, better designated as *blocking*, the participants are divided into two or more strata and then randomly assigned to treatments within strata (blocks). For the answer sheet example, if a researcher were carrying out the experiment to test the effect of the new way (the experimental treatment) versus the old way (the control treatment) for a sample that consisted of 10 males and 20 females and wanted to be sure that there were proper proportions of males and females in each of the treatments, he or she would randomly assign 5 of the males to the experimental treatment and the other 5 males to the control treatment, and would randomly assign 10 of the females to the experimental treatment and the other 10 females to the control treatment. That would permit testing for the "main effect" of treatment, the "main effect" of sex, and the sex-by-treatment "interaction effect."

There is occasional confusion in the literature between random *assignment* and random *sampling*. The purpose of the former, as noted earlier, is pre-experimental equivalence of treatment groups so that post-treatment differences can be attributed to the effects that are caused by the treatments themselves. Thus the objective is one of *internal validity*. The purpose of the latter, however, is to be able to generalize from a sample to the population from which the sample has been drawn. That objective is concerned with *external validity*.

Traditional approaches to sample-to-population inferences such as the *t*-test have often been incorrectly used to analyze the data for true experiments in which random assignment has been employed but random sampling has not; that is, the participants have been selected by some sort of convenience sampling procedure. The appropriate inferential statistical technique for such studies is a randomization test (sometimes called a permutation test). The actual difference between treatment groups is compared with all possible differences that could have been obtained under different randomizations.

Ideally, a study should possess both generalizability and causality; that is, researchers would like to be able to employ both random sampling and random assignment. For nonexperimental research, a researcher might have random sampling but not random assignment (e.g., because there are no "treatments" to "assign" subjects to), in which case there would be a statistical basis for generalizability to the population but an insufficient basis for assessing causality. Or a researcher may not have the luxury of either random sampling or random assignment. That does not necessarily mean that the researcher should not carry out the study and report its results. But it does mean that the researcher will be restricted to the use of descriptive statistics only, with any sort of generalizable or causal interpretation being necessarily subjective.

Although the objectives of random assignment and random sampling are different, researchers can use the same random-number "tables" for accomplishing both. There are computerized routines available on the Internet that provide, at no charge, a random sample of n numbers out of N numbers (where n is less than or equal to N) and/or a random assignment of N numbers into subsets of $n_1, n_2, \ldots, n_k$, where the sum of those subscripted n's is equal to N, and k is the number of treatments. For an example, see the Research Randomizer Web site, which offers a quick way to generate random numbers or assign participants to experimental conditions.

Thomas R. Knapp

See also Experimental Design; External Validity; Interaction Effect; Internal Validity; Main Effect; Random; Random Sampling

Further Readings

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). London: Chapman & Hall/CRC.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *American Statistician*, *52* (2), 127–132.

Orwin, R. G., & Boruch, R. F. (1982). RRT Meets RDD: Statistical strategies for assuring response privacy in telephone surveys. *Public Opinion Quarterly*, *46* (4), 560–571.

RAND Corporation. (2001). *A million random digits with 100,000 normal deviates*. Glencoe, IL: Free Press. (Original work published 1955)

Research Randomizer: http:www.randomizer.org

Random-Digit Dialing (RDD)

Random-digit dialing (RDD) refers to a set of techniques for drawing a sample of households from the frame or set of telephone numbers. The telephone number is the sampling unit that is the link to the household and its members. While the specific sampling techniques employed to draw RDD samples have changed over time, the three most influential RDD techniques are briefly described in this entry. RDD is distinguished from other telephone sampling methods because RDD selects the sample from the frame of telephone numbers, whereas the other methods select from lists of numbers in directories or commercial lists. The ability to sample all telephone households, not just those households on a list, is one reason for the popularity of RDD sampling. This entry ends by discussing two issues that often arise in RDD sampling: selecting RDD samples for local areas and selecting persons within the household to interview.

The basic RDD approach is simple. Using information on the structure of the telephone numbering scheme used in North America, the set of all numbers that could be assigned to households is identified. These numbers are randomly sampled, often with equal probability. The techniques for randomly selecting the numbers are defined by the specific RDD sampling scheme. The sampled numbers are dialed, and those that are residential are the sampled households. In many RDD surveys only members in a certain age range or of a specific sex are eligible to be interviewed. The data collected from the sample are used to make estimates or inferences. The estimates usually refer to all households, even though households without telephones are not covered.

RDD surveys became popular by the late 1970s, and the technique soon became the predominant method of sampling households in the United States. The popularity of RDD coincided with evidence that a large proportion of households lived in telephone households and could be sampled or covered in RDD surveys. By 1980, over 90% of adults lived in households with landline telephones and the percentage continued to grow slightly, to around 97% by 2000.

An important advantage of RDD sampling that helped to spur its acceptance is the relatively low cost of sampling and conducting surveys by telephone as compared to face-to-face surveys. An RDD sample is likely to cost less than 20% of the costs of an area probability sample that has the same precision. These great cost advantages fueled the early acceptance of RDD in commercial surveys. It also is the feature that makes it possible to conduct surveys of rare population groups with probability samples. For example, the Centers for Disease Control and Prevention conduct a survey to monitor childhood vaccinations in each state and large metropolitan area, even though they sample only households with children between 19 and 35 months old, that is, approximately 1 in 50 U.S. households. The cost of this type of survey by face-to-face methods would be astronomical.

Despite these advantages, the utility of RDD surveys has been questioned in recent years. One concern is the decrease in response rates in telephone surveys. Response rates in RDD surveys have always been lower than in face-to-face surveys but higher than in mail surveys conducted without extensive follow-up mailings. However, as response rates have continued to decline in RDD surveys, the potential for nonresponse bias in estimates has increased. An important reason for the lower response rates is that technological advances have made it easier for households to identify and avoid telephone calls from unrecognized telephone numbers.

Another challenge to RDD samples is the emergence of wireless or cell telephones as a major method of communication in the United States. As a larger proportion of households drop their landlines and subscribe only to wireless telephone service, the potential for noncoverage bias in estimates from RDD surveys increases because RDD surveys traditionally exclude wireless telephones. The percentage of U.S. households that have only wireless service has increased from less than 1% around 2000 to 20% by the end of 2007. A related problem for RDD samples is that even in some households that have retained a landline, household members may rely on their wireless telephones to such an extent that interviewers may not be able to contact the household members on their landlines. By the end of 2007, approximately 10% of U.S. households fell into this category.

Other issues confronting RDD surveys include the introduction of Voice over Internet Protocol, number portability, and national do-not-call lists. The effects of most of these have not yet become clear. If RDD is to retain its role as a premier method for sampling households, these topics must be addressed. Fortunately, the history of RDD sampling suggests that survey researchers are innovative, and further developments in this tradition are possible. The key methods for selecting RDD samples (described next) exemplify this innovation.

RDD Sampling Methods

The three most important RDD sampling techniques are (a) the original RDD method, (b) Mitofsky-Waksberg RDD sampling, and (c) list-assisted RDD sampling. Because all three of these RDD methods depend on the structure of telephone numbers in the United States, a review of the U.S. telephone number plan is given before summarizing the sampling techniques.

Each telephone number in the United States has 10 digits; the first 3 are called the area code, the next 3 are the prefix, and the last 4 are the suffix. The first 8 digits are sometimes collectively called "100-banks" because they define sets of 100 possible telephone numbers with the same first 8 digits. Area codes have specific geographic boundaries, although these boundaries are not necessarily consistent with geopolitical boundaries other than state lines. In recent years, multiple area codes have been assigned that cover portions of the same geography in densely populated areas. Prefixes within an area code cannot be directly associated with a smaller geographic area, but households within a prefix traditionally have tended to be clustered spatially. Prefixes within an area code are assigned a designation that identifies the type of service (e.g., regular household service, paging, cellular) associated with these numbers. Although these designations are generally accurate, a small percentage of numbers may be for services not consistent with the designation for the prefix.

The Original RDD Sampling Method

The original RDD sampling technique was formally introduced in 1964. The approach involved identifying area code–prefix combinations assigned for regular household telephones, and then appending a random 4-digit number to create a sampled telephone number. This process was repeated until the desired sample size was achieved. In this scheme, each number has an equal probability of selection. Households, on the other hand, have unequal probabilities of selection because they may have more than one telephone line in the household; thus, the household could be sampled on any of these numbers. But this inequality is easily remedied by either weighting adjustments or further subsampling, as is often done in case-control RDD samples. In the early days, the area code–prefix combinations were identified from local telephone directories, but later eligible combinations became available nationally from computerized files. In this RDD method, all possible telephone numbers, not just those listed in telephone directories, could be sampled. Furthermore, by taking advantage of the area code–prefix combination, the efficiency of sampling (i.e., contacting a residence with the dialed number) increased from less than 1% to over 20%. Alternatives to further improve the efficiency of the calling were offered, such as the add-a-digit method, but most of these methods suffered from other problems that limited their acceptance.

Mitofsky-Waksberg (MW) RDD Sampling

The next major advance was named the Mitofsky-Waksberg (MW) RDD sampling technique after its two developers, Warren Mitofsky and Joseph Waksberg. The method was developed after it was observed that residential numbers tended to be clustered within 100-banks. More specifically, a large portion of the eligible 100-banks had no residential numbers assigned, whereas in those with any residential numbers, over 30% of the numbers were residential. This was due largely to the ways local telephone companies assigned the numbers to customers. The MW scheme takes advantage of this clustering to improve calling efficiency using an equal probability, two-stage sampling scheme. The first stage begins with the random selection of 100-banks from the frame of eligible area code–prefixes. Two random digits are appended to form 10-digit telephone numbers that are dialed. If a sampled number is not residential, then the 100-bank is not sampled in the first stage. If the number is residential, then the 100-bank is sampled, and a set of additional numbers within the 100-bank are sampled to achieve a fixed number of residences within the 100-bank that are sampled. This scheme is very elegant theoretically because 100-banks are sampled with probability exactly proportional to the number of residential numbers in the bank even though the number of residential numbers is unknown at the time of sampling. All households have the same probability of selection, with the same caveat given previously concerning multiple lines within the household. The biggest advantage of the MW method is its efficiency: About 60% to 65% of the numbers sampled in the second stage are residential. This was a marked improvement over the original RDD sampling. Due to its efficiency, the MW technique became the main method of RDD sampling for nearly two decades.

List-Assisted Sampling

Despite the popularity of the MW sampling technique, many alternatives were suggested because the method had some implementation and operational difficulties and the sample was clustered (the 100-banks are the first-stage clusters). The clustering reduced the statistical efficiency of the sample. In the early 1990s alternative methods of stratified RDD sampling were being explored, and technological developments allowed easy and inexpensive access to data to support these schemes. The most recent method of RDD sampling, list-assisted RDD sampling, was spawned from this environment and was quickly adopted by practitioners. In list-assisted RDD sampling all 100-banks in eligible area code–prefix combinations are classified by the number of directory-listed telephone numbers in the bank. A sample of 100-banks with x or more listed telephone numbers is selected (when $x = 1$ this is called the listed stratum). The other stratum (called the zero-listed stratum when $x = 0$) may be sampled at a lower rate or may not be sampled at all because it contains so few residential telephone numbers. Commercial firms developed systems to implement this method of sampling inexpensively. When research showed that excluding the zero-listed stratum resulted in minimal noncoverage bias, list-assisted sampling from only the listed stratum became the standard approach.

The list-assisted method is an equal probability sampling method that avoids both the operational and statistical disadvantages of the MW method. The calling efficiency of the list-assisted technique is approximately equal to that of the MW method. The efficiency has decreased over time as there has been less clustering of residential numbers in 100-banks, but the reduced clustering also affects the MW method. The lower percentage of residential numbers is also less of a problem because many organizations use predictive dialing to lower the labor costs associated with nonresidential numbers. Even those organizations that do not use predictive dialing themselves often take advantage of telephone purging methods provided by sampling vendors. After purging, about 70% of the sample numbers remaining are residential. By the mid-1990s, list-assisted RDD sampling, excluding the zero-listed stratum, had displaced the MW method as the primary RDD sampling technique in the United States.

Sampling Issues With RDD

The methods for RDD sampling could be used to conduct a household survey in which any adult in the household reports for the entire household. However, typical RDD surveys require sampling one individual or more to be interviewed, such as any adult or persons with a particular characteristic (e.g., all women over 55 years). Many methods for sampling one adult from a household are in the literature, but techniques for other sampling requirements are less frequently discussed. Frequently, the household is screened to determine if any person in the household is eligible. If there are any eligible members, then all the eligible persons or a sample of them is selected.

Another issue arises when estimates are needed for a local area, such as a city or county, rather than the entire nation or state. In local areas, it may not be possible to cleanly identify area code–prefix combinations that serve the local area because the assignments do not follow political boundaries. Sampling for local areas became more difficult with the introduction of number portability, which often allows households to keep their telephone numbers when they move out of their local area. One sampling approach for local areas is to include only those combinations that are almost entirely within the area, accepting the loss in coverage from households outside these combinations and number portability. A more precise, but expensive, approach is to include all the combinations that serve the local area and screen out those households that are sampled but do not live in the area. Other choices are possible, but all result in some compromise between coverage of the target population and cost of data collection.

J. Michael Brick

See also Add-a-Digit Sampling; Cluster Sample; List-Assisted Sampling; Mitofsky-Waksberg Sampling; Number Portability; Telephone Surveys; Voice over Internet Protocol (VoIP) and the Virtual Computer-Assisted Telephone Interviewing (CATI) Facility

Further Readings

Blumberg, S., Luke, J., & Cynamon, M. (2006). Telephone coverage and health survey estimates: Evaluating the need for concern about wireless substitution. *American Journal of Public Health*, *96*, 926–931.

Brick, J. M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, *59*, 218–235.

Cooper, S. L. (1964). Random sampling by telephone: An improved method. *Journal of Marketing Research*, *1*, 45–48.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Nathan, G. (2001). Telesurvey methodologies for household surveys—a review and some thoughts for the future. *Survey Methodology*, *27*(1), 7–31.

Tucker, C., Lepkowski, J. M., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly*, *66*, 321–338.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, *73*, 40–46.

Random Error

Random error refers to the fact that any survey measure taken over and over again may well be different (by some small amount) upon each measure, merely due to chance measurement imprecision. Thus, compared to the true value of the measure for a given respondent, the observed value will be on the high side some of the time and on the low side other times. This deviation of the observed value from the true value often is signified by e in the following formula:

$$X = T + e,$$

where X is the observed value and T is the true value. In theory, over many similar measures of the same variable taken from the same respondent, the average of the observed values will equal the true value. That is, the random error in measuring the variable of interest will deviate from the true value so that the sum of those deviations is expected to be zero. Random error is not consistent error and thus, by definition, it is not a part of bias, which is consistent (i.e., directional) *nonrandom* error.

The concept of random error serves as the basis for statistical inference. By measuring the size of the random error, statistical tests can be applied to determine the confidence with which a statistical inference can be drawn. Every sample statistic will have its associated random error, which typically is expressed as its standard error. Researchers have some control over the magnitude of the size of random error by increasing the sample size or using a sampling design that is more precise. The trade-off is that these techniques raise survey costs.

Paul J. Lavrakas

See also Bias; Sample Size; Standard Error; Statistic; True Value; Variance

Further Readings

Agresti, A., & Finley, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice Hall.
Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Randomized Response

Researchers who study sensitive topics are often confronted with a higher refusal rate and often obtain more socially desirable answers. To tackle these problems, Stanley L. Warner introduced the randomized response technique (RRT). This is an interview method that guarantees total privacy and therefore, in theory, can overcome the reluctance of respondents to reveal sensitive or probably harmful information. Warner's original method used a randomization device (usually colored beads, coins, or dice) to direct respondents to answer one out of two statements, such as:

A: *I am a communist*. (A: selected with probability p)

B: *I am not a communist*. (not-A: selected with probability $1 - p$)

Without revealing to the interviewer which statement was selected by the dice, the respondent answers true or not true according to whether or not he or she is a communist. Elementary probability theory can be used to get a bias-free estimate ($\hat{\pi}$) of the population probability of A (being a communist).

Profits and Costs of Using Randomized Response Techniques

The advantage of randomized response is that more valid population estimates of sensitive behavior can be derived. The cost is that randomized response is less efficient than a direct question. The randomization procedure increases the sampling variance, which makes it necessary to use larger samples. For example, Warner's method typically needs as many as 10 times the number of respondents to be just as powerful as a direct question design.

Extra costs are also associated with the increased complexity of the randomized response question. When randomized response is used, respondents have to understand and follow the instructions, in addition to understanding and formulating a response to the question itself. This introduces a new source of error, namely, misunderstanding the RRT procedures, cheating on this procedure, or both.

The advantage of using a randomized response technique outweighs the extra costs only when the population estimates become significantly more valid than estimates obtained from direct question–answer designs. Meta-analysis has shown that the advantage of using randomized response outweighs the disadvantage of having to use a larger sample if the social sensitivity of a research topic is large.

Other Randomized Response Techniques

Since Warner's original idea, many adaptations and refinements have been developed. First, methods were developed to improve the power of the design. As previously noted, compared to a direct question, Warner's method typically needs much larger samples. Variations on Warner's original method have been developed that have a larger statistical power. These methods include the *unrelated question technique* and the *forced response technique*. In addition, randomized response methods have been developed that are easier for the respondent, such as *Kuk's card method* and the *item count technique*.

The Unrelated Question Technique

The unrelated question technique uses a randomization device to direct the respondent to one of two unrelated questions, the sensitive question and an innocuous question, for example:

Do you have (sensitive attribute A)?

Do you have (the nonsensitive attribute B)?

The unrelated question technique is most efficient if the prevalence of the nonsensitive attribute B is known. In that case only one sample is needed to compute the unbiased estimate for the sensitive attribute A. An example of such a nonsensitive question is for instance: *Is/was your mother's birthday in July?* In this case, a sample of about twice the normal sample size is needed for the question to be as efficient as a direct question.

A disadvantage of the unrelated question method is that it is difficult to find nonsensitive questions with known prevalence in the population. When the prevalence of the nonsensitive question is not known, a second sample is needed to estimate the prevalence, and the advantage of the unrelated question technique over Warner's original method is lost.

The Forced Response Technique

In the forced response technique, only the sensitive question is asked, and the randomization device specifies the direction of the answer. For example, a pair of dice could be used, and respondents are requested to answer the sensitive question truthfully with "Yes" or "No" if the total roll of these dice is between 5 and 10. When the roll of the dice is 2, 3, or 4, the respondent is forced to answer "Yes," independent of the true score on the sensitive question. When the dice roll 11 or 12, the respondent is forced to answer "No," regardless of the true score on the question. Again, the meaning of an individual respondent's "Yes" or "No" answer is nil (and unknown to the researcher), but knowledge of the proportion of forced responses allows the researcher to estimate the prevalence of sensitive attribute A in the population.

The forced response technique is as efficient as the unrelated question method with known prevalence of the nonsensitive question. As a result, the sample has to be only about twice as large compared to direct question samples.

Kuk's Card Method

Kuk developed a randomized response method that avoids requiring the respondent to give direct answers like "Yes" or "No." A respondent receives two decks of cards and a sensitive question. One pack of cards is named the Yes-pack, to be used when the respondent should respond with "Yes" on sensitive attribute A. The respondent picks a card from the Yes-pack and names the color of the card: red or black. When the respondent's true answer is "No," the card of the other pack is read. The major disadvantage of Kuk's method is that it is very inefficient; in fact, statistically it is equivalent to Warner's original method.

The Item Count Technique

The item count technique, also known as the *unmatched count technique* and the *list-experiment technique*, was developed to avoid the extra cognitive burden on the respondent and the distrust that sometimes is associated with the use of a randomizer. The method is compellingly simple. There are two lists with activities: One list contains only innocent activities, and the other list contains one additional activity, the sensitive activity of interest. Respondents are randomly assigned to one of the lists. Each respondent is asked to report the number of activities he or she has engaged in but not the name of the activities. The difference in mean activities between the first and the second list is an estimate for the prevalence of the sensitive attribute.

The advantages of this method are the absolute anonymity of the respondent, the lower cognitive burden for the respondents, and compared to the direct question method, the power does not decrease. However, if the sensitive activity is rare, its variance will be very small, compared to the total variance. This will lessen the reliability of the estimated prevalence of the sensitive activity.

Analysis of Randomized Response Data

For all randomized response techniques, straightforward analysis techniques exist to estimate the prevalence of the sensitive behavior and the corresponding sampling variance. For all techniques except the item count technique, it is also possible to use a modified logistic regression analysis to estimate the relationships of the sensitive behavior with respondent characteristics. Again, the cost of randomized response techniques is that the standard errors of the regression weights will be larger than with ordinary logistic regression of data obtained by direct questioning. In addition, special software is needed to carry out this analysis; standard statistical packages cannot be used.

Joop Hox and Gerty Lensvelt-Mulders

See also Disclosure Limitation; List-Experiment Technique; Privacy; Sensitive Topics; Social Desirability

Further Readings

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 185–210). New York: Wiley.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. Beverly Hills, CA: Sage.

Hox, J. J., & Lensvelt-Mulders, G. (2004). Randomized response analysis in Mplus. *Structural Equation Modeling*, *11*(4), 615–620.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods & Research*, *33*(3), 319–348.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69.

Random Order

Random order refers to the randomization of the order in which questions appear in a questionnaire. The purpose is to overcome a type of measurement error known as context effects. This randomization is most often done using a computer program that controls a computer-assisted interview (CAI) being conducted in person, over the phone, or self-administered. Prior to the common use of CAI in survey research, sets of questions were randomized using modified Kish tables that were generated prior to the beginning of the survey and printed on labels indicating to the interviewers what order the questions were to be asked. The order changed randomly for each label to be printed. The labels were pasted next to the question sets in the paper version of the questionnaire. The questions in the set were asked in the random order of the numbers on the label.

There are many issues that impact the response process when a question is presented to a respondent. The items immediately preceding a specific question often have no consequence. Concrete questions addressing facts such as demographic characteristics or behaviors are less likely to be affected by the context effect of the previous questions. However, questions requiring an attitude or opinion can more readily be influenced by the issues addressed in previous questions. Questionnaires are often designed so that a series of topically related attitude questions are presented one after another. In this situation, the likelihood of earlier questions affecting the interpretation of, and responses to, latter questions increases. Presenting the questions in this type of set in a random order minimizes the likelihood of this type of measurement error.

A simple example of this is the set of questions regarding abortion that have been included for many years in the General Social Survey:

Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if:

1. *The woman's own health is seriously endangered by the pregnancy.*
2. *There is a strong chance of a serious defect in the baby.*
3. *She became pregnant as a result of rape.*
4. *The family has a very low income and cannot afford any more children.*
5. *She is not married and does not want to marry the man.*
6. *She is married and does not want any more children.*
7. *The woman wants it for any reason.*

Each of these questions is about abortion, but the scenarios elicit a variety of potential emotions, value judgments, and other unpredictable issues that might influence a person's response to successive questions. There is strong reason to believe that asking the last question (number 7) first each time is likely to result in different percentages of positive and negative responses to questions 1 through 6 than if it were always asked after all the other questions.

This problem is overcome by randomizing the order of the questions in the set for each interview. Current software used in personal (CAPI), self- (CASI), and telephone interviewing (CATI) or in Internet surveying makes implementation of this process quite easy by automatically randomizing question order at presentation but reordering the data for ease of analysis. The resulting aggregate results present a more balanced view than might otherwise be achieved if the questions were always asked in the same order.

James Wolf

See also Context Effect; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); General Social Survey (GSS); Measurement Error; Questionnaire-Related Error

Further Readings

Davis, J. A., Smith, T. W., & Marsden, P. V. (Eds.). National Opinion Research Center (Producer). (2005). General Social Surveys, 1972–2004 [Cumulative data file] (Publication No. ICPSR04295-v2). Storrs: University of Connecticut, Roper Center for Public Opinion Research. Retrieved May 1, 2008, from http://www.ropercenter.uconn.edu/data_access/data/datasets/general_social_survey.html

Random Sampling

Random sampling refers to a variety of selection techniques in which sample members are selected by chance, but with a known probability of selection. Most social science, business, and agricultural surveys rely on random sampling techniques for the selection of survey participants or sample units, where the sample units may be persons, establishments, land points, or other units for analysis. Random sampling is a critical element to the overall survey research design.

This entry first addresses some terminological considerations. Second, it discusses two main components of random sampling: randomness and known probabilities of selection. Third, it briefly describes specific types of random samples, including simple random sampling (with and without replacement), systematic sampling, and stratification, with mention of other complex designs. The final section touches on inference, which is the reason that random sampling is preferred in scientific surveys.

Terminological Considerations

Some authors, such as William G. Cochran, use the term *random sampling* to refer specifically to simple random sampling. Other texts use the term *random sampling* to describe the broader class of probability sampling. For this reason, authors such as Leslie Kish generally avoid the term *random sampling*. In this entry, random sampling is used in the latter context, referring to the broader class of probability sampling.

Critical Elements

The two critical elements of random sampling are randomness and known probabilities of selection.

Randomness

The first critical element in random sampling is the element of randomness. Ideally, all members in the survey's target population have a non-zero chance of selection.

In describing what random sampling is, it is helpful to highlight what it is not. The sample is not pre-determined. Nor is a random sample selected by expert judgment. Random sampling does not imply that the sampling is haphazard. Furthermore, random sampling is not convenience sampling, in which the interviewers take respondents that are easiest to obtain.

The element of randomness is applied to the process scientifically. That is, there is a method, usually mechanical, to the selection process that is rigorously followed. The precise method may rely on random number generators or tables of random numbers. By following the scientific process, the probabilities of selection are known and preserved.

Random number generators and tables of random numbers are not truly random, of course, but the process needs to be random enough. This is especially important in litigious contexts. Bruce D. McCullough and Wendy Rotz have tested the random number generators available in various data tools and statistical packages.

Known Probabilities of Selection

The probabilities of selection are important for the theory that enables researchers to estimate sampling error. Because a sample is a subset of the target population and not a census (complete enumeration), estimates derived from sample responses will rarely match the target population values exactly. The variable difference between the sample estimate and the population value is sampling error. (Nonsampling errors, such as inaccurate frames of the target population and imprecise measures of the questionnaire items, affect both surveys and censuses. Nonsampling errors are not covered in this entry.)

Having a randomly selected sample with known probabilities of selection enables the researcher to estimate the sampling error. That is, the researcher can use the sample to make inferences for the target population and to estimate the precision of the sample-based estimates.

The probabilities of selection may enter into the estimation process in another way, as well. Because the probabilities of selection for members of a random sample are known, the sample responses can be appropriately weighted (if the probabilities are different, as in complex sampling designs) to yield improved estimates for the target population. The weights are a function of the probabilities of selection, which are not known precisely for purposive, convenience, or other nonprobability samples.

Sample Designs

Simple Random Sampling

The most familiar type of random sampling is *simple random sampling*. Simple random sampling may be with or without replacement. The drawing of names out of a hat and the selection of official lotto numbers are examples of simple random sampling without replacement. In simple random sampling, all members of the population have the same probability of selection.

The selection is said to be *without replacement* if a member of the population cannot be selected more than once in the same sample. Usually the sample members are selected sequentially. Each selection is made from the population excluding those already selected. Therefore, the sample draws are not independent.

Simple random sampling is said to be *with replacement* if each selected sample member is replaced into the available population pool for subsequent draws. In practice, sampling with replacement is not as common as sampling without replacement.

An easy way of selecting a simple random sample of size n without replacement is to use a random number generator to assign a random number to each member of the population in the frame or population list, sort the frame by the random number, and select the first n units in the randomly ordered list.

Systematic Sampling

Another random sampling design is *systematic sampling*, in which the population is ordered, and every kth unit is selected. Once a random starting point is selected, the rest of the sample is determined; however, the randomness is in the selection of the starting point. In other words, a single sample is randomly selected from a set of k possible samples. (If k is not an integer, more samples are possible.)

Complex Designs

In practice, simple random sampling and systematic sampling are rarely used alone for large surveys, but are often used in combination with other design features that make for a more complex design. Complex random sampling designs tend to have smaller sampling error or lower cost, or both. The complex random sampling designs may incorporate elements that resemble purposive sampling, such as stratification. Double sampling, cluster sampling, multi-stage sampling, and sampling with probability proportional to size are other examples of complex probability or random sampling.

Stratified Sampling

Stratification involves dividing the target population into groupings, or strata, of relatively homogeneous members and selecting a random sample independently within each stratum. For example, a list of individuals may be divided into strata based on age and gender, or the population of businesses may be divided into strata based on geography, size, and industry. The variables for stratification need to be available for all population members and presumably related to the survey response variables or the propensity to respond. Because members within a stratum tend to be more alike, the emphasis is on selecting a broad sample across strata rather than extensively within each stratum. By selecting members from each stratum, the sample will capture much of the diversity in the population more efficiently than if the sample were selected purely randomly. Stratification also enables the researcher to sample from the strata at different rates if, for example, the researcher wants estimates for individual strata as well as for the population as a whole.

Inference

The survey design includes both the sampling technique and the corresponding estimators for inferences. With scientific random sampling and known probabilities, mathematical formulas exist for making inferences about the target population and for estimating the sampling error attributed to the inferences. Confidence intervals, tests of hypotheses, and other statistics are possible with random sampling and estimates of sampling error. While an expert may judiciously select a sample that is a good representation of the target population on some measure, a purposive sample of this sort cannot, by itself, be used to estimate the precision of the sample-based estimates because no such mathematical formulas are possible. Neither can the sampling error be estimated from a convenience sample.

Under simple random sampling, the distribution of the sample mean often approximates the normal distribution, where the variance decreases with sample size. That is, the sample mean is a good estimator for the population mean, and the error associated with the sample-based estimate is smaller for larger samples. This result is based on the central limit theorem.

Alternatively, in some circumstances, especially when sample sizes are small, the distribution of sample statistics may approximate Poisson, hypergeometric, or other distributions. The approximate distribution is what enables the researcher to make inferences about the population based on sample estimates.

Complex probability designs may have more complex mathematical forms for statistical inference and may require specialized software to properly handle the estimation. Some methods of variance estimation developed for complex designs may be less stable in some circumstances, an issue to be aware of as the field moves toward smaller samples and estimation using subsets of the sample. Nevertheless, inference is possible because complex designs share the underlying concepts and theory of random selection with known probabilities.

Rachel Harter

See also Cluster Sample; Cochran, W. G.; Complex Sample Surveys; Convenience Sampling; Kish, Leslie; Multi-Stage Sampling; Nonprobability Sample; Nonsampling Error; Probability of Selection; Probability Proportional to Size (PPS) Sampling; Probability Sample; Purposive Sampling; Random; Sampling Error; Sampling Without Replacement; Simple Random Sample; Strata; Stratified Sampling; Systematic Sampling

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Liu, Y., Batcher, M., & Rotz, W. (2001, August). Application of the hypergeometric distribution in a special case of rare events. *Proceedings of the Section on Survey Research Methods* (pp. 941–954). Alexandria, VA: American Statistical Association.

Marsaglia, G. (1995). *The Marsaglia Random Number CDROM including the Diehard Battery of Tests of Randomness*. Retrieved May 1, 2008, from http://www.stat.fsu.edu/pub/diehard/

McCullough, B. D. (1998). Assessing the reliability of statistical software: Part I. *American Statistician*, *52*(4), 358–366.

McCullough, B. D. (1998). Assessing the reliability of statistical software: Part II. *American Statistician*, *53*(2), 149–159.

National Institute of Standards and Technology. (1980). *NMS Minimal BASIC Test Program for random generators* (Special Publication No. 500-70/1 and No. 500-70/2). Gaithersburg, MD: Author.

National Institute of Standards and Technology. (2001). *A statistical test suite for random and pseudorandom number generators for cryptographic applications* (Special Publication No. 800-22). Gaithersburg, MD: Author.

Rotz, W., Falk, E., & Joshee, A. (2004, August). A comparison of random number generators used in business—2004 update. *Proceedings of the Section on Business and Economic Statistics* (pp. 1316–1319). Alexandria, VA: American Statistical Association.

Rotz, W., Falk, E., Wood, D., & Mulrow, J. (2001, August). A comparison of random number generators used in business. *Proceedings of the Annual Meeting of the American Statistical Association, Section on Computational Statistics*. Alexandria, VA: American Statistical Association.

Random Start

The term *random start* has two separate meanings in survey research—one related to questionnaire item order and one related to sampling. This entry discusses both meanings.

Random Start in Questionnaire Construction

In terms of questionnaire item order, a random start means that a series of similar items (such as a set of agree–disagree questions) is administered in a way that the first question in the series that is asked of any one respondent is randomly chosen to "start" the series, and then the order of the subsequent questions is not randomized. For example, if Q23 through Q28 (six items in all) made up a set of attitude questions that use an agree–disagree response scale, by using a random start a researcher will have one sixth of the respondents asked Q23 first, one sixth asked Q24 first, one sixth Q25 first, and so on. All respondents are asked all the questions in the series, and each question follows its numerical predecessor, except when it is the first item asked in the series; for example, Q26 always follows Q25, except when Q26 is the first question asked (then Q25 is asked last as the random start order of the set would be Q26-Q27-Q28-Q23-Q24-Q25). The purpose of using a random start is to help control for possible item order effects, so that not all respondents will be asked the questions in the exact same order.

A random start set of questions differs from a random order set of questions in that the latter randomizes all questions within the set in all possible ways and then randomly assigns a respondent to one of the possible randomized orders. Random start questions are more readily analyzed and interpreted than random order questions, especially when there are more than a few questions in the set. Thus, for example, even with a set of 10 items using a random start deployment, a researcher with an overall sample size of 1,000 would have 100 respondents for each of the 10 random start orders in which the items would be asked. That would allow the researcher to investigate for order effects in how the random start series of questions were answered. In the case of a set of 10 items that were asked in a true random order, the permutations would be so great that the researcher could not investigate order effects with confidence unless the sample size were enormous (far greater than almost any survey would ever gather).

Random Start in Sampling

In terms of sampling, a random start refers to randomly selecting the first element in a systematic sampling procedure in order to avoid sampling error. Most forms of commonly used statistical analysis are based on the assumption that the data were selected using a procedure that allows for the calculation of a nonzero selection probability for each element in the sample frame. The least complex probability sampling procedure is simple random sampling, but this is not always an easy procedure to implement because it essentially requires that a separate random number be generated for each element to be selected for inclusion in the sample. Applied field work rarely presents the researcher with machine-readable lists of every eligible element of the population under study. Systematic sampling is often the preferred alternative.

A systematic sample procedure is a two-step process, and it uses a random start. The first step is to determine the sampling interval (k) by dividing the population size (N) by the desired sample size (n). The second step is to determine the random start by selecting a random number between 1 and k. Note that the selection probability is zero for elements that are not exact multiples of k unless a random start is used. Use of the random start and a properly calculated interval allows for a nonzero selection probability to be calculated for each element in the population. Like simple random sampling, the systematic sampling approach usually results in a sample that is generalizable to the population without the need for weights. Unlike simple random sampling, the systematic approach requires only one random number to be generated rather than one for each element selected.

The success of using a random start and set interval selection thereafter rests on a properly prepared sample frame. Care should be taken to ensure that the population sample frame to be used is either randomized or stratified sufficiently so that the interval selected via the random start will not match cycles in the sample frame order. For example, records reflecting daily aggregate data will have a 7-day cycle; thus, an interval that is a multiple of 7 will select only 1 day of the week. Similar cycles (months, semesters, seasons, etc.) must be considered in relation to the random start and interval in order to avoid sampling error through systematic exclusion.

Another issue to consider is when the population size (N) is not an integral multiple of the interval (k). In this case, the final interval has fewer elements than the rest. If the random start is greater than the size of the last interval, the target sample size (n) will not be achieved. This is not usually an issue in survey research, because of the larger sample sizes that surveys most often use ($n > 100$). However, there are several easily implemented solutions to this problem when smaller sample sizes are warranted.

James Wolf

See also Elements; Probability Sample; Question Order Effects; Random Order; Sampling Error; Sampling Frame; Simple Random Sample; Systematic Sampling

Further Readings

Kish, L. (1995). Systematic sampling: Stratification techniques. In *Survey sampling* (pp. 113–147). New York: Wiley.

RANKED-SET SAMPLING (RSS)

The most basic sampling technique to use, when collecting data from a population for a sample survey, is that of simple random sampling (SRS). Ranked-set sampling (RSS) is an alternative probability sampling technique to SRS. While the items in an simple random sample might or might not be mutually independent (depending on whether sampling is with or without replacement), it is always the case that such a sample is designed so that each measured observation can be viewed as "representative" of the underlying population. Even with this probabilistic guarantee, however, there is still the possibility that a given simple random sample, just by mere chance, might not represent the underlying population well. That has led statisticians to consider a variety of ways to guard against obtaining unrepresentative samples.

Approaches for Ensuring Representative Samples

One way to better ensure representative samples is to put additional structure on the sampling design. Some examples of this approach include *stratified element sampling* and *stratified cluster sampling*. Sampling designs can become increasingly complex when the population of interest is large and diverse, such as when the goal is to collect data on a national sample. Although the primary goal is to select sampling units that are representative of the underlying population, a secondary, but often just as important, goal is to minimize the costs associated with collecting the data, that is, both the cost of measurement and that associated with obtaining the sample units.

An alternative cost-effective approach to obtaining more representative sample observations from a population is that of RSS. The RSS technique uses additional information about potential sample units as an aid in choosing which of the units should actually be measured on the variable(s) of interest. In this way information about all of the units selected for potential measurement is used to guide the selection of the specific units to be measured. It is this additional information that enables RSS techniques to generally outperform analogous SRS techniques when both involve the same number of measured observations.

Example of Ranked-Set Sampling

To provide a concrete illustration of how RSS is conducted, we consider the setting where our goal is to estimate the unknown population mean, $\bar{X}$, using the information in n measured observations. RSS takes advantage of available information from additional potential sample units to enable us to measure selected units that are, collectively, more representative of the population of interest. The net result of RSS is a set of measurements that are more likely to span the range of values in the population than can be guaranteed from SRS. Following is a more precise description of this process.

Suppose we wish to obtain a ranked-set sample of k measured observations. First, an initial simple random sample of k units from the population is selected and rank-ordered on the attribute of interest. This ranking can result from a variety of mechanisms, including expert opinion (called *judgment ranking*), visual comparisons, or the use of easy-to-obtain auxiliary variables; it cannot, however, involve actual measurements of the attribute of interest on the sample units. The unit that is judged to be the smallest in this ranking is included as the first item in the ranked-set sample, and the attribute of interest is formally measured on this unit.

Denote the measurement obtained from the smallest item in the rank ordered set of k units by $x^*_{(1)}$. The remaining $k-1$ unmeasured units in the first random sample are not considered further in the selection of this ranked-set sample or eventual inference about the population. The sole purpose of these other $k-1$ units is to help select an item for measurement that represents the smaller values in the population.

Next, a second independent random sample of size k is selected from the population and judgment ranked without formal measurement on the attribute of interest. This time we select the item judged to be the second smallest of the k units in this second random sample and include it in our ranked-set sample for measurement of the attribute of interest. This second measured observation is denoted by $x^*_{(2)}$.

From a third independent random sample we select the unit judgment ranked to be the third smallest, $x^*_{(3)}$, for measurement and inclusion in the ranked-set sample. This process is continued until we have selected the unit judgment ranked to be the largest of the k units in the k^{th} random sample, denoted by $x^*_{(k)}$, for measurement and inclusion in the ranked-set sample.

The entire process described above is referred to as a *cycle*, and the number of observations in each random sample, k, is called the *set size*. Thus, to complete a single ranked set cycle, we judgment rank k independent random samples of size k each, involving a total of k^2 sample units to obtain k measured observations $x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(k)}$. These k observations represent a *balanced ranked-set sample* with set size k, where *balanced* refers to the fact that we have collected one judgment order statistic for each of the ranks $I = 1, \ldots, k$. To obtain a ranked-set sample with a desired total number of measured observations $n = km$, we repeat the entire cycle process m independent times, yielding the data $x^*_{(1)j}, \ldots, x^*_{(k)j}$, for $j = 1, \ldots, m$. The RSS estimator of the population mean $\bar{X}$ is

$$\bar{x}_{RSS} = \frac{1}{km} \sum\nolimits_{j=1}^{m} \sum\nolimits_{i=1}^{k} x^*_{(i)j}.$$

We illustrate the application of RSS methodology to a small sample of data from the National Health and Nutrition Examination Survey III (NHANES III). The NHANES III data set contains various body measurements and information on health-related variables for the respondents. For our example we consider the two variables body mass index (BMI) and thigh circumference (TC). We demonstrate how to use TC values to collect a structured ranked-set sample to estimate the mean BMI for the population of nonpregnant adults in the NHANES III data set.

We provide the details for estimating the mean BMI using a ranked-set sample of size $n = 6$ with a set size $m = 6$ and a single cycle ($k = 1$).

Step 1: Collect six independent random samples of size $m = 6$ potential sample units from the NHANES III data set. We obtain the following $n = 6$ subsets (random samples) of size $m = 6$ (BMI, TC) data pairs each. (For an actual ranked-set sample the BMI values would not be available until after the 6 subjects for the sample had been selected. Then we would only need to measure the BMI values for those 6 subjects, not for all 36 potential subjects. We include the BMI values for all 36 subjects here, however, for use in our comparison of the RSS estimates with competitors based on similar-sized simple random samples.)

Subset 1: (19.4, 43.7) (26.5, 52.1) (24.3, 48.3)
(34.0, 49.1) (23.9, 50.6) (25.8, 50.3)
Subset 2: (26.0, 49.0) (26.0, 46.3) (23.2, 50.7)
(32.6, 56.0) (24.3, 45.9) (22.9, 50.7)
Subset 3: (26.3, 49.5) (34.9, 56.4) (20.9, 46.2)
(32.3, 46.9) (28.3, 56.0) (25.0, 50.8)
Subset 4: (29.7, 53.5) (27.3, 50.8) (17.5, 42.9)
(27.6, 50.4) (24.9, 45.9) (22.5, 42.3)
Subset 5: (28.4, 52.1) (24.1, 49.1) (25.0, 52.5)
(28.1, 48.7) (24.6, 50.4) (23.7, 52.0)
Subset 6: (25.6, 42.5) (28.7, 55.6) (22.4, 46.4)
(21.4, 50.0) (23.3, 49.8) (22.9, 44.3)

Step 2: Rank order the pairs on the basis of their TC values (in bold) from smallest to largest, separately in each of the six subsets. The ranked pairs for our data are the following:

Subset 1: (19.4, **43.7**) (24.3, **48.3**) (34.0, **49.1**)
(25.8, **50.3**) (23.9, **50.6**) (26.5, **52.1**)
Subset 2: (24.3, **45.9**) (26.0, **46.3**) (26.0, **49.0**)
(23.2, **50.7**) (22.9, **50.7**) (32.6, **56.0**)
Subset 3: (20.9, **46.2**) (32.3, **46.9**) (26.3, **49.5**)
(25.0, **50.8**) (28.3, **56.0**) (34.9, **56.4**)
Subset 4: (22.5, **42.3**) (17.5, **42.9**) (24.9, **45.9**)
(27.6, **50.4**) (27.3, **50.8**) (29.7, **53.5**)
Subset 5: (28.1, **48.7**) (24.1, **49.1**) (24.6, **50.4**)
(23.7, **52.0**) (28.4, **52.1**) (25.0, **52.5**)
Subset 6: (25.6, **42.5**) (22.9, **44.3**) (22.4, **46.4**)
(23.3, **49.8**) (21.4, **50.0**) (28.7, **55.6**)

Step 3: For the i^{th} subset, "measure" and record the BMI associated with the i^{th} ordered TC value, $i = 1, \ldots, 6$. The resulting six recorded BMI values are 19.4, 26.0, 26.3, 27.6, 28.4, and 28.7.

Step 4: Compute the average for the six recorded BMI values obtained in Step 3. This is the RSS estimate of the population mean BMI. For our single cycle of ranked-set sample data, the estimate of the mean BMI is

$$\bar{x}_{BMI,RSS} = (19.4 + 26.0 + 26.3 + 27.6 + 28.4 + 28.7)/6 = 26.067.$$

Additional structure for RSS, compared to SRS, is achieved from the use of all 36 TC measurements to aid in the rankings and the eventual decision about which BMI measurements to obtain. It is this structure inherent in ranked-set sample data resulting from the use of additional concomitant information that leads to the improved properties of the RSS estimator.

Advantages of Random-Set Sampling

Statistical theory shows that the sample average from a ranked-set sample consisting of n measured observations is a more efficient estimator of the population mean than a simple random sample with the same number of measured observations. Both sample averages are unbiased estimators of the population mean, but the variance of the ranked-set sample average is never any greater than the variance of the simple random sample average. Thus the use of RSS leads to increased precision or, equivalently, to reduced sample sizes with the same precision relative to SRS.

This improvement from using an RSS approach to data collection as opposed to the standard SRS approach extends to many other settings. In particular, a ranked-set sample can be taken within strata, or a ranked-set sample of clusters could be selected. In fact, using RSS instead of SRS at the final stage in any sampling scheme will generally lead to estimators with smaller variances.

Douglas A. Wolfe and Elizabeth A. Stasny

See also Cluster Sample; National Health and Nutrition Examination Survey; Simple Random Sample; Stratified Sampling; Unbiased Statistic

Further Readings

Chen, Z., Bai, Z. D., & Sinha, B. K. (2004). *Ranked set sampling: Theory and applications*. New York: Springer.

Wolfe, D. A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science*, *19*, 636–643.

Ranking

There are a variety of formats that can be used in asking survey questions, from items that require a simple "yes" or "no" response to other types of forced-choice items, rating scales, and multi-part items in which respondents' opinions are determined through a series of questions. Ranking is a question response format used when a researcher is interested in establishing some type of priority among a set of objects, whether they be policies, attributes, organizations, individuals, or some other topic or property of interest.

For example, if city officials were interested in whether citizens thought it was most important to improve city services in the area of police protection, fire protection, garbage and trash collection, road maintenance and repair, or parks and recreation, one method they could use would be to use an open-ended question that asked respondents, *What service that the city provides do you think is most important to improve in the next 12 months?* The percentage of respondents that mentioned each of these services could then be used as a measure of priority for service improvement. Another method for determining this would be to ask a series of rating questions in which respondents were asked whether it was very important, somewhat important, not too important, or not at all important for the city to improve services in each of these five areas, and the option that received the highest percentage of "very important" responses or had the highest average rating would have priority for service improvement. A third way to accomplish this would be through a ranking question, for example: *The city provides a number of services, including police protection, fire protection, garbage and trash collection, road maintenance and repair, or parks and recreation. Of these, which do you think is most important for the city to improve in the next 12 months?* After respondents selected their "most important" service, they would then be asked, *And which service is next most important for the city to* *improve?* three times to identify their second, third, and fourth priority for service improvement, with the service not selected being ranked fifth by default.

Because ranking requires making choices from among a series of options, this format is generally thought to be more difficult (i.e., to create a greater cognitive burden) for respondents than other response formats. This is particularly true for telephone surveys in which the ranking of five items is considered to be the maximum number of objects that the average respondent accurately can rank. Rankings can be used more readily in face-to-face interviews or in a self-administered questionnaire when visual cues (e.g., show cards) can be used to assist respondents in remembering the options and making adjustments in their priority ordering, though there is a limit on the number of ranked items that should be displayed even with these methods of data collection.

When the research question requires a larger number of items to be ranked, several alternatives have been developed. One method is to have respondents rank only those items at each end of their preferences, leaving the middle preferences unranked. For example, if 15 items were to be ranked in terms of their perceived threat to the security of the United States, the three biggest threats could be selected (and these three ranked, if desired, to produce first, second, and third rankings of threat). Similarly, the three least threatening items could be selected (and ranked), producing differentiation at the "biggest" and "least" ends of the scale, but less differentiation in the middle.

Another method for obtaining rankings is through a series of paired comparisons. In using paired comparisons, a respondent considers each object in comparison with each of the other alternatives, one at a time. If there is an ordered quality to these objects (i.e., object A is preferred to B, B is preferred to C, and A is preferred to C), then paired comparisons provide a fairly robust method for ranking objects. In practice, however, respondents' choices are not always consistent, making the priority ranking of objects less clear. The number of paired comparisons that can be made is also limited to some extent, in that the number of comparisons that have to be made increases geometrically with the number of options that need ranking; for example, with 6 objects, 15 paired comparisons are required.

As with any decision about the format of the questions to be used in a survey, the determination of when to use rankings and the number of objects to be ranked should be guided by the research question of interest. In many research situations, the results produced by rankings, by a series of rating questions, or even by open-ended questions will produce similar results. Rankings have certain advantages, as well as disadvantages, relative to other formats; the substantive question being investigated should determine when their use is most appropriate.

Robert W. Oldendick

See also Forced Choice; Open-Ended Question; Paired Comparisons; Rating; Respondent Burden; Response Alternatives; Show Cards

Further Readings

Fowler, F. J., Jr. (1993). *Survey research methods* (2nd ed.). Newbury Park, CA: Sage.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys*. Thousand Oaks, CA: Sage.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Rare Populations

A rare population is generally defined as a small proportion of a total population that possesses one or more specific characteristics. Examples include billionaires; people with a certain illness, such as gall bladder cancer; or employees in a highly technical occupation. Although the literature offers no precise definition of *rare* or *small* in this context, researchers have proposed proportions of .10 or less to identify rare populations. (When this proportion is larger, standard sampling techniques can usually be used efficiently.) In addition, sampling frames are often nonexistent or incomplete for most rare populations. Although researchers can use convenience sampling (e.g., snowball samples) to study rare populations, most efforts in this area have focused on probability sampling of rare populations. The costs and benefits of the various approaches can be difficult to define a priori and depend on the type and size of the rare population.

Sampling Strategies

Generally, sampling frames for the total population do not contain information identifying members of the rare population; if they do, then the sampling process is trivial. If not, then screening must be used to identify members of the group. With screening, members of the rare population are identified at the beginning of the interview process. However, the costs of screening can often exceed the costs of interviewing, especially if the rare population is only a very small proportion of the sampling frame.

There are several ways in which screening costs can be reduced. Mail questionnaires can be used to identify members of the rare population if the sampling frame has correct address information; for in-person interviews, the use of telephone screening can reduce interviewer costs. If the rare population is geographically clustered, for example, in certain states or urban areas, screening based on clusters becomes less costly.

Besides cost, another potential drawback to the screening approach is response errors of commission and omission (i.e., false positives and false negatives, respectively) during the screening process, especially if many different questions must be correctly answered in order to identify a member of the rare population. Using less stringent criteria during screening to identify members is one approach to reducing response errors, because misclassified members (false positives) can be excluded after the full interview is complete. The main problem here is devising a screening process that avoids erroneously classifying members of the rare population as nonmembers (false negatives).

Whereas for many rare populations it is impossible to derive a complete sampling frame, one or more incomplete lists may be available. For example, hospital or pharmacy records may identify some, but not all, members of a population with a specific rare disease. A dual frame approach could then be used, in which the partial list is combined with screening of the total population to reduce screening costs. Alternatively, the partial list could be used with a cluster sampling approach to identify areas where members of the rare population are located.

In some situations multiplicity or network sampling can be used to locate and interview members of a rare population. Typically, a member of a household is interviewed and queried as to whether other members of the household or close relatives are members of a special population (although respondents also can be asked about neighbors or members of other social groups to which they belong). Occasionally, the researcher may need only an estimate of the size of the rare population; however, in most cases the researcher wishes to interview members of the rare population. If so, accurate contact information may be difficult to obtain from the respondent. In general, accuracy of reporting depends on the relationship between the respondent and the member of the rare population and the visibility of whatever characteristic(s) defines the rare population. For example, World War II veteran status is generally well known and thus visible to network members, whereas certain illnesses may not be. Researchers must use special weighting when deriving estimates from network samples, because the probability of selection can vary by the size of the network. Costs may be higher than with typical screening methods, especially if addresses are sought.

Special location samples are useful when the rare population is defined by an activity (e.g., spearfishing) or because the behavior that defines the rare population may not always be known by members of a network (e.g., married men who have sex with other men). With this approach, the researcher identifies specific locations where members of the rare population may congregate, usually through extensive field work, and potential members of the rare population are then approached at varying times throughout the day. Known as *time-space sampling* or *venue sampling*, this area of research has emphasized randomly selecting locations and times in an effort to yield a probability sample. Although costly, this approach can be used to study rare populations that might defy other sampling efforts.

Finally, adaptive sampling procedures have often been used to study rare populations, but their main drawback is the lack of a probability sample that can be used to make generalizations about the entire rare population. *Adaptive sampling* refers to a group of similar sampling procedures that "adapt" to the sampling situation based on information gathered during the sampling process. For example, members of a rare population may be asked to identify other members of a rare population; once identified, these new members are added to the sampling frame. *Snowball sampling* is one example of an adaptive sampling approach. Several scholars have developed estimation methods for adaptive samples that allow generalization. Generally, these methods depend on assumptions about the probability distribution generating the sample, although recent work has attempted to relax this restriction.

Stephen R. Porter

See also Adaptive Sampling; Cluster Sample; Convenience Sampling; Dual-Frame Sampling; Errors of Commission; Errors of Omission; Multiplicity Sampling; Network Sampling; Screening; Snowball Sampling

Further Readings

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A (General)*, *149*(1), 65–82.

Kish, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, *7*(3), 339–347.

Muhib, F. B., Lin, L. S., Stueve, A., Miller, R., Ford, W. L., Johnson, W. D., et al. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports*, *116*(Suppl. 1), 216–222.

Sudman, S., & Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, *12*, 401–429.

Sudman, S., Sirken, M., & Cowan, C. D. (1988). Sampling rare and elusive populations. *Science*, *240*(4855), 991–996.

Thompson, S. K., & Collins, L. M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, *68*(Suppl. 1), S57–S67.

Rating

In constructing a survey questionnaire there are a number of response formats that can be used in asking respondents about their views on various topics. One of the most commonly used formats in survey questionnaires is the rating, or rating scale.

As the name implies, ratings are regularly used in surveys to evaluate an object along some dimension, such as performance or satisfaction. In a typical rating question, respondents are asked to make judgments along a scale varying between two extremes, such as from “very good” to “very poor,” or “extremely positive” to “extremely negative,” and the like. The following illustrates a common rating scale question: *How would you rate the job (PERSON’s NAME) is doing as governor . . . Would you say she is doing an excellent job, a good job, a fair job, a poor job, or a very poor job?* As another example, the item, *Some people don’t pay much attention to political campaigns. How about you . . . would you say that you have been extremely interested, very interested, not too interested, or not at all interested in the political campaigns so far this year?* is a rating scale in which respondents are asked to assess their interest in campaigns among choices ranging from “extremely” to “not at all.” A third example is the item, *Now I’d like to ask you to rate the city as a place to live and work on a scale of 0 to 10, with 10 being the highest rating you could give and 0 being the lowest*.

A rating question provides the topic to be evaluated, the dimension along which it is to be evaluated, and the scale of the evaluation. As the previously mentioned examples illustrate, rating questions vary widely in terms of the dimensions on which objects are evaluated (job performance, interest in political campaigns, place to live and work), the scales used (“excellent” to “very poor,” “extremely interested” to “not at all interested,” 0 to 10), and the number of response options (5, 4, and 11, respectively).

The various components of a rating question affect the results produced by such items. For example, in evaluating presidential performance, different—although very likely similar—results would be obtained if the item asked the respondents to assess such performance as *excellent, good, fair, poor,* or *very poor* than if it asked them whether they *strongly approved, somewhat approved, neither approved nor disapproved, somewhat disapproved*, or *strongly disapproved* of the president’s performance. Even a seemingly minor change in the options provided—for example, changing “excellent” to “very good”—can yield different results and sometimes these differences are not negligible.

Another important factor in rating questions is the number of response options. It is generally considered that five options are the most respondents can understand without some type of visual aid, although some surveys use more, such as the “rate on a scale of 0 to 10” example described previously. Related to this is whether respondents should be provided with an odd or even number of response categories. Rating questions with an odd number of categories provide respondents with a true “middle” option, whereas items with an even number of categories force respondents who feel “in the middle” to lean in one direction or the other. The order in which the options are presented, from “most positive” to “least positive” or vice versa, can also make a difference in the results obtained. For any of these considerations, there are no technically right or wrong ways in which to ask a rating question. Each of the decisions made in developing an item should be guided by the research question of interest.

The mode used for survey data collection—face-to-face, telephone, mail, Web, or some other self-administered format—also can affect the types of rating questions that can be asked. Because visual aids cannot be presented over the telephone (at least not without a videophone), such surveys are more limited in the number of response options that can be provided. As a result, telephone surveys often employ "unfolding" questions. For example, in being asked their opinion on a particular policy, respondents may first be asked to rate whether they *favor, oppose*, or are *neutral*. Those who initially favor the option are then asked whether they favor it strongly or favor it only somewhat; those who initially oppose are asked to rate if they oppose it strongly or only somewhat; and initially neutral respondents are asked if they lean toward favoring or opposing it. The end result is a 7-point scale with the options *favor strongly*; *favor somewhat*; *lean toward favoring*; *neutral, lean toward neither*; *lean toward opposing*; *oppose somewhat*; and *oppose strongly*. Providing respondents with visual aids in face-to-face or self-administered surveys allows or more options in the number of response alternatives, such as a "feeling thermometer," which is based on the concept of a thermometer and typically asks respondents to rate some object from 0° (very cold or unfavorable) to 100° (very warm or favorable).

Robert W. Oldendick

See also Feeling Thermometer; Forced Choice; Likert Scale; Mode of Data Collection; Precoded Question; Ranking; Response Alternatives; Unfolding Question

Further Readings

Converse, J., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Newbury Park, CA: Sage.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys*. Thousand Oaks, CA: Sage.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Ratio Measure

Ratio measure refers to the highest (most complex) level of measurement that a variable can possess. The properties of a variable that is a ratio measure are the following: (a) Each value can be treated as a unique category (as in a nominal measure); (b) different values have order of magnitude, such as greater than or less than or equal to (as in an ordinal measure); (c) basic mathematical procedures can be conducted with the values, such as addition and division (as with an interval measure); and (d) the variable can take on the value of zero. An example of a ratio measure is someone's annual income.

A ratio measure may be expressed as either a fraction or percentage; in addition, a ratio measure may be written as two numbers separated by a colon. For example, if Susan earns $40 per hour and John earns $20 per hour, then the fraction of Susan's pay that John earns can be expressed as 2/4, the percentage of Susan's pay that John earns as 50%, and the ratio of John's pay to Susan's as 1:2.

There are many situations in which a ratio measure is more appropriate than a total or mean or other descriptive statistic when reporting the results of a survey. Ratio measures are utilized in the business world, in health care, in banking, and in government, as well as in other applications.

In the business world, there are a number of applications involving ratio measures. For example, suppose that one purpose of a survey of a sample of businesses is to assess liquidity. A commonly used measure of liquidity is the current ratio measure; this ratio measure is an important indicator of whether the company can pay its bills and remain in business. It is calculated by dividing a company's total current assets by total current liabilities for the most recently reported quarter. Thus, if a business has total current assets of $450,000 and total current liabilities of $200,000, then its liquidity ratio measure would be $450,000/$200,000 = 2.25. However, in the sample, a number of businesses are sampled and each reports its total current assets and total current liabilities. To determine the liquidity of the sample of businesses, one would calculate the total of the reported assets of all of the sampled businesses and divide by the total of the reported liabilities of the same businesses. That is, if the total assets of the sample equal $10 billion and the total liabilities equal $5 billion, then the cost ratio measure of the sample equals $10,000,000,000/ $5,000,000,000 = 2.0. One would then generalize to the population from which the sample of businesses was drawn.

Carla R. Scanlan

See also Interval Measure; Level of Measurement; Nominal Measure; Ordinal Measure

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Raw Data

The term *raw data* is used most commonly to refer to information that is gathered for a research study before that information has been transformed or analyzed in any way. The term can apply to the data as soon as they are gathered or after they have been cleaned, but not in any way further transformed or analyzed. The challenge for most researchers who collect and analyze data is to extract useful information from the raw data they start with.

First, it is important to note that *data* is a plural noun (*datum* is the singular noun). Second, all collections of raw data are, by definition, incomplete. Some data are uncollected because of a variety of constraints. Lack of time and money are common reasons researchers fail to collect all available raw data to answer a given research question. Increasingly, however, data collection has become easier and more efficient in many ways, and more and more of the data that are intended to be collected are collected. (Of note, a counter trend in survey research has led to a decrease in the amount of intended data that are able to be collected. This is the trend for sampled respondents to be more reluctant to cooperate with survey requests.) Absent the ability to process this additional raw data, merely having more data may not increase knowledge, but enhanced computing power has produced an ability to process more data at little or no additional cost.

Third, the expression *raw data* implies a level of processing that suggests that these data cannot provide useful information without further effort. *Raw* suggests the data have not yet been summarized or analyzed in a way so as to "release" the information for which it was collected.

Fourth, raw data may be collected in alternative representational schemes that affect how the researcher thinks about processing the data. Perhaps the easiest way to think about this is to consider the variety of symbolic formats used in computer processing of information. Data are commonly represented in binary, hexadecimal, or decimal number systems, or as ASCII (American Standard Code for Information Interchange) text. Data organization is also relevant: Information may be organized as bits, nibbles, bytes, words, or other such units.

In some cases, one researcher's processed data may be another's raw data. The methodology by which data are collected and the level of aggregation of collected data play important roles in defining information as raw data to be analyzed. Consider, for example, the collection of survey responses to a series of five questions designed to measure attitudes toward abortion. At one level, raw data are the individual responses to each of the five questions. These data might be used by a survey researcher to establish the consistency of each respondent's answers to the individual survey questions. Such efforts might produce a second level of data in which a "scale" is created to represent the responses to the series of questions, with one scale score assigned to each individual. A third level of data aggregation might be reached by summarizing the individual scale scores for subsamples or groups of the individuals interviewed. A fourth level of aggregation of these data would be to release data on what Benjamin Page and Robert Shapiro call "collective public opinion," by reporting the average scale score for the entire sample—usually in comparison to another sample taken from the same population at a different point in time or to a sample drawn from a different population at the same point in time.

Even more critical than the methodology used to collect data is the research question to be answered or the theory to be tested. Data must be identified and collected with a purpose. Research questions motivate the collection of some subset of all possible data and the exclusion of other information.

Raw data are not the exclusive bailiwick of quantitative (social) science. Scientists of many traditions collect data. The raw data of qualitative social scientists may include taped or transcribed interviews, impressions and judgments, and archival material of many types and varieties. Each of these pieces of information is combined with others to allow judgments and inferences.

Raw data are similarly collected and analyzed by businesses and governments for their own purposes. Corporations collect basic information down to the level of the individual transaction to assess the capacity and profitability of their company. Newspapers have common rules for translating raw data into news. Information is collected from multiple sources, and "facts" are cross-checked and verified to assure that accurate information has been collected and reasonable inferences drawn by the reporter. Similarly, analysts at the Central Intelligence Agency work with raw data, from mechanical telemetry to in-field human intelligence, to produce estimates of the capacities of foreign governments.

Calling data "raw" does not absolve its user of responsibility for assessing its quality. All data should be subject to a variety of assessments, including judgment as to the degree of error contained in the data relevant to the purpose for which the data were collected. Raw materials should be evaluated for accuracy and completeness. If raw materials are "sampled" from a larger universe, their representativeness should be considered. Assessments include establishing the reliability and validity of data as well as justifications for discarding some of the data and consideration of the problem of missing data.

John P. McIver

See also Missing Data; Quality Control; Reliability; Total Survey Error; Validity

Further Readings

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.
Diesing, P. (1991). *How does social science work?* Pittsburgh, PA: University of Pittsburgh Press.
Hyde, R. (1996). *The art of assembly language*. San Francisco: No Starch Press.
Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.
Page, B. I., & Shapiro, R. Y. (1992). *The rational public*. Chicago: University of Chicago Press.

Reactivity

Reactivity occurs when the subject of the study (e.g., survey respondent) is affected either by the instruments of the study or the individuals conducting the study in a way that changes whatever is being measured. In survey research, the term *reactivity* applies when the individual's response is influenced by some part of the survey instrument (e.g., an item on a questionnaire); the interviewer; the survey organization sponsor conducting the study, or both; or the environment where the survey is taking place. For example, the respondent may respond positively or negatively based on the interviewer's reactions to the answer. A smile, nod, frown, or laugh may alter how the subject chooses to respond to subsequent questions. Deliberate or accidental, the actions of the interviewer administering the survey instrument may affect the subject's response.

A second instance of reactivity is when the subject reacts to the instrument itself. An example of this is respondents who respond to questions based on how they wish to see themselves or the environment in which they are completing the survey, rather than answering accurately. The same subjects may answer the same questions differently depending on where the survey was completed—for instance, in a homeless shelter or a country club. Respondents especially may be sensitive to their sincere answers that are widely opposed or criticized. Whether deliberate or unintentional, the actions of the interviewer, the environment of the survey, or the survey instrument itself may affect the accuracy of the subject's response.

Reactivity is undesirable in social science research because it decreases the validity and veracity of the project's results. Unless the subject's response to the environment, the survey instrument, or the experimenter is the focus of the study, these stimuli may introduce nonreplicable and confusing effects into a research project. Loss of validity results from miscalculating the impact of parts of the project unrelated to the research question. The research project then drives the response from the subject, and the survey instrument reflects the faults of the study instead of the accurate answers of the respondent.

To avoid the problem of reactivity, a researcher begins with a sound research design. There are three ways to reduce reactivity in a survey. First, when designing the project, a researcher must be aware of how different individuals may react to different aspects of the research experience. Depending on the group under study, cultural, economic, and other social differences may yield unwanted results based on some parts of the questionnaire. Whenever possible, the research should also guard against environmental factors influencing the respondent. Second, a researcher needs to review the work of other analysts and investigate problems in other scholarly research projects. By taking note of the faults of other studies, researchers can avoid or remove similar difficulties in their own work. Finally, each interviewer must be well trained and monitored. The research design should include interviewer guidelines for behavior and demeanor when they are conducting the survey. Controlling these three areas of the study will reduce the likelihood of reactivity effects and thereby strengthen the external validity of the research project.

Ryan Gibb

See also External Validity; Interviewer Monitoring; Interviewer-Related Error; Interviewer Training; Measurement Error; Research Design

Further Readings

Berg, B. (1998). *Qualitative research methods for the social sciences*. Boston: Allyn & Bacon.
Manheim, J., Rich, R., Willnat, L., & Brians, C. (2006). *Empirical political analysis: Research methods in political science*. New York: Pearson Education.
Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Recency Effect

A recency effect is one type of response order effect, whereby the order in which response options are offered to respondents affects the distribution of responses. Recency effects occur when response options are more likely to be chosen when presented at the end of a list of response options than when presented at the beginning. In contrast, primacy effects occur when response options are more likely to be chosen when presented at the beginning of a list of response options than when presented at the end. Response order effects are typically measured by presenting response options in different orders to different groups of respondents. For example, if half of the respondents in a survey are asked, *Which of the following is the most important problem facing the country today: the economy or lack of morality?* and the other half are asked, *Which of the following is the most important problem facing the country today: lack of morality or the economy?* a recency effect would be observed if a greater proportion of respondents chose the economy in response to the second question than in response to the first.

In the many studies of response order effects in questions with categorical response options, primacy effects were observed in some cases, recency effects were observed in others, and no significant response order effect was observed in others. One explanation for the mixture of findings comes from satisficing theory. Survey researchers hope respondents will answer questions carefully and thoughtfully, but respondents may not always be able or motivated to do so. Instead, they may shift their response strategies to minimize effort while providing a satisfactory response to the survey question (i.e., known as *satisficing*). One such strategy involves choosing the first response option that seems reasonable, and this strategy is believed to be responsible for response order effects.

When response options are presented visually, most respondents likely begin by considering the option presented first, then the second option, and so on. So if respondents choose the first reasonable response option they consider, primacy effects are likely to occur. But when response options are presented orally as happens in interviewer-administered surveys, respondents cannot think much about the first option they hear, because presentation of the second option interrupts this thinking. Similar interference occurs until after the last response option is heard, and at that point the last response option is likely to be the most salient and the focus of respondents' thoughts. People may also be most likely to remember the last response options in a long list of response options. So if respondents choose the first reasonable response option they consider, recency effects will occur. Consistent with this logic, mostly primacy effects have appeared in past studies that involved visual presentation, and mostly recency effects have occurred under oral presentation conditions.

In addition to mode of presentation, satisficing theory posits that response order effects depend on three factors: (1) the respondent's ability, (2) the respondent's motivation, and (3) the cognitive difficulty of optimizing inherent in the question. Respondents with greater ability and motivation are less likely to satisfice. Satisficing is also more likely when a question's stem or response choices are especially difficult to comprehend, when a question demands an especially difficult search of memory to retrieve needed information, when the integration of retrieved information into a summary judgment is especially difficult, or when translation of the summary judgment onto the response alternatives is especially difficult. Thus, recency effects in questions with orally presented, categorical response options are likely to be strongest among respondents low in ability and motivation, and for questions that are more difficult.

Allyson Holbrook

See also Primacy Effect; Response Alternatives; Response Order Effects; Satisficing

Further Readings

Bishop, G., & Smith, A. (2001). Response-order effects and the early Gallup split-ballots. *Public Opinion Quarterly*, *65*, 479–505.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, *67*, 205–211.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research*, *20*, 60–103.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58–88.

Payne, S. L. (1949). Case study in question complexity. *Public Opinion Quarterly*, *13*, 653–658.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.

Recoded Variable

Analysis with survey research data often first requires recoding of variables. Recode is a term used to describe the process of making changes to the values of a variable. Rarely can researchers proceed directly to data analysis after receiving or compiling a "raw" data set. The values of the variables to be used in the analysis usually have to be changed first. The reasons for recoding a variable are many. There may be errors in the original coding of the data that must be corrected. A frequency distribution run or descriptive statistics on a variable can help the researcher identify possible errors that must be corrected with recodes. For example, if a variable measuring respondent's party identification has values of 1 for Democrat, 3 for Independent, 5 for Republican, and 7 for Other, and a frequency distribution shows values of either 2, 4, or 6, then these incorrect values can be recoded to their correct value by going back to the original survey instrument if it is available. More likely these incorrect values will have to be recoded to a "missing value" status. Computer-assisted data collection makes it impossible to check the correct value, as there is no paper trail, but makes it less likely that incorrect values will be entered in the first place.

Data are usually collected to get as much information as possible, but these data often must be recoded to yield more interpretable results. For example, respondents may be asked to report their date of birth in a survey. Recoding these values to age (in years) lets the researcher interpret the variable in a more intuitive and useful way. Furthermore, if the researcher wants to present the age variable in a frequency distribution table, then the interval measure of age can be recoded into an ordinal measure. For example, respondents between the ages of 18 and 24 can have their age value recoded to category 1, respondents between 25 and 34 can be recoded to category 2, respondents between 35 and 44 can be recoded to category 3, and so on. As interval- and ratio-level data can always be recoded into nominal- or ordinal-level data but nominal- and ordinal-level data cannot be recoded into interval-level data, it is always better to collect data at the interval or ratio level if possible.

Recoding is also necessary if some values of a variable are recorded inconveniently for analysis. For example, surveys typically code responses of "Don't know" and "Refused" as 8 and 9. The researcher needs to recode these values so they are recognized as missing values by the computer program being used for statistical analyses. Leaving these values unchanged will yield inaccurate and misleading results. Another reason to recode a variable is to transform the values of a variable (e.g., a log transformation) to tease out the true nature of the relationship between two variables. Also, if missing values are assigned a number in the original coding, the researcher needs to recode these values to missing before analyzing the data. As a final precaution, it is advisable to always run a frequency or descriptive statistics on the recoded variable to make certain that the desired recodes were achieved.

Charles Tien

See also Coding; Computer-Assisted Telephone Interviewing (CATI); Don't Knows (DKs); Frequency Distribution; Interval Measure; Nominal Measure; Ordinal Measure; Ratio Measure; Raw Data; Variable

Further Readings

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Nagler, J. (1995). Coding style and good computing practices. *The Political Methodologist*, *6*, 2–8.

Nardi, P. M. (2003). *Doing survey research: A guide to quantitative methods*. Boston: Allyn & Bacon.

RECONTACT

The term *recontact* has more than one meaning in survey research. It is used to refer to the multiple attempts that often must be made to make contact with a sampled element, such as a household or person, in order to gain cooperation and gather data and achieve good response rates. Recontacting hard-to-reach respondents is the most effective way to reduce survey nonresponse that is due to noncontact. This holds whether the mode of contact is via multiple mailings, multiple emails, multiple telephone calls, or multiple in-person visits.

The other meaning of *recontact* in survey research refers to the efforts that are made to validate completed interviews soon after they have been completed, by having a supervisory staff member recontact the respondent to validate that the interview was in fact completed and that it was completed accurately. These recontacts often are conducted via telephone even when the original interview was completed by an interviewer face-to-face with the respondent. Recontacts of this type are reasoned to reduce the likelihood of interviewer falsification, assuming that interviewers know that such quality assurance efforts are carried out by survey management.

Paul J. Lavrakas

See also Falsification; Noncontacts; Nonresponse; Quality Control; Response Rates; Validation

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

RECORD CHECK

A record check study is one in which the validity of survey self-reports is assessed by comparing them to evidence in organizational records. Record check studies have been employed to evaluate the quality of data from existing surveys and also to test the relative quality of data from different prospective survey designs. Such investigations are often viewed as the gold standard for judging the validity of self-report information on behavior and experiences.

Record checks are infrequently conducted because of the limited availability of record information, as well as the cost and effort of obtaining and matching record and survey data for the same individuals. Although there are many potential sources of records, in practice it can be difficult to gain access to information that gives insight into the quality of survey data. The cooperation of record-holding organizations is required and, unfortunately, often hard to obtain. Records must be collected and preserved in a manner that permits investigators to use the information to make valid inferences. Organizations assemble records for their own aims, not for the purpose of validating survey data. Thus, information that survey investigators deem important may be missing or poorly maintained in records. Records may be organized in such a way that they require substantial transformation before they can be employed in a record check study. Finally, records are not kept on many things about which survey investigators are interested; record check studies are therefore confined to those matters in which human behavior comes in contact with bureaucratic systems. Other important behaviors, as well as attitudes, perceptions, and other mental constructs cannot be examined through the record check technique.

These points notwithstanding, it is clear that evidence from record check studies has been influential in understanding reporting errors and in designing important surveys. Record checks have been conducted in a number of contexts—for example, health, labor, voting, and victimization. In research done in connection with the National Health Interview Survey, medical records were employed to check the quality of reporting of doctor visits, hospitalizations, and morbidity. Records were obtained from physicians and hospitals detailing the date of medical encounters, along with the reasons for the visits. Respondents were then interviewed about their health experiences, and the survey reports were matched against the records. Comparisons revealed, for example, whether respondents reported a recorded medical contact or not. A result of one of these studies was that the reference period for the question on doctor visits was set at 2 weeks before the interview because comparison of survey reports with visit records showed that accuracy of recall was sharply lower for longer reference periods.

Another important series of record checks involved the American National Election Study. For several post-election waves of this biennial survey, respondent reports of voting were checked against public voting records. In this case, respondents were asked about their voting first and then interviewers attempted to verify the reports by examining voting records in county clerk offices. These comparisons suggested that there is an overreporting bias in self-reported voting. More people said that they voted than actually did. Since these studies were conducted, there have been many papers written on the possible causes and consequences of vote overreporting. In addition, the American National Election Study question used to measure voting was modified in an attempt to reduce the hypothesized "social desirability" bias attached to self-report of voting. Finally, there have been investigations of whether statistical models predicting voting differed in their predictive power if the analysis was confined to only "valid" votes—those reported that matched the official records—or also included the "invalid" votes. The findings suggest that the models performed equivalently in these cases.

In the development and the redesign of the National Crime Victimization Survey, police records were employed to validate reports of victimization. People who had reported crimes to the police were interviewed and asked if they had been victimized. The comparison of record information and self-reports suggested that victimization is underreported, particularly (as in the case of medical events) as the time between the police report and the survey interview increased and for less serious crimes. These findings influenced the designers of the National Crime Victimization Survey to set a 6-month reference period for asking questions about victimization and to employ a variety of cues to assist respondents in recalling victimization experiences.

The research on the National Health Interview Survey, the American National Election Survey, and the National Crime Victimization Survey was preceded by a famous study conducted by Hugh Perry and Helen Crossley in Denver, Colorado, in the late 1940s. This study asked respondents about, for example, voting, library card ownership, and traffic offenses. The self-reports were matched against public records, and the results suggested that voting and library card ownership were overreported, whereas traffic offenses were underreported.

Combining the findings from all of these record check studies, general inferences have been made about factors affecting the validity of self-reports. First, social desirability is thought to affect the reporting of certain sorts of behavior: The more a behavior is thought to be meritorious (e.g., voting), the more it is likely to be overreported, whereas the more a behavior is thought to be undesirable (e.g., traffic offenses), the more it is apt to be underreported. Second, the elapsed time between an event and the survey is believed to affect the quality of self-reports: The longer the interval between the event (e.g., a doctor visit) and the interview, the less likely the event is to be recalled or reported accurately in time. Third, the importance or salience of an event is believed to affect the quality of reports. For example, more serious victimizations are more likely to be reported accurately than less serious ones, and hospitalizations that lasted a number of days are more likely to be reported accurately than overnight visits. Thus, record check studies have influenced greatly how survey practitioners think about possible reporting error and the means to reduce it.

Although these lessons are important, the picture presented by record check studies is not so clear that survey researchers can treat the inferences from them as "laws" of survey reporting. The basic assumption of record check research—that the records are generally free of error—is certainly questionable. For example, an investigation of police records in Chicago in the 1980s led to the suspicion that the records were being systematically altered to make it appear that crimes had been solved. Ironically, investigators in that case employed a survey of possible victims to see if what they said matched the records—but this time, they treated the victims' responses as "the truth." Further, voting records are not always correct. It may also happen that records and self-reports are both "true," even if they do not match. This is because official records capture human events differently from the way people experience them. For example, records of people receiving a governmental housing subsidy may not be reported by the recipients simply because they do not know that the rent they pay is subsidized. In sum, records, like self-reports, are fallible sources of information, and meticulous care must be taken in record check studies to see to it that proper inferences are drawn.

Peter V. Miller

See also National Election Studies (NES); National Health Interview Survey (NHIS); Reference Period; Respondent-Related Error; Reverse Record Check; Self-Reported Measure; Social Desirability; Telescoping

Further Readings

Cannell, C. F., Fisher, G., & Bakker, T. (1965). *Reporting of hospitalization in the Health Interview Survey*. Vital and Health Statistics, Series 2, No. 6. Hyattsville, MD: National Center for Health Statistics.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389–437). San Francisco: Jossey-Bass.

Miller, P. V., & Groves, R. M. (1985). Matching survey responses to official records: An exploration of validity in victimization reporting. *Public Opinion Quarterly*, *49*, 366–380.

Parry, H., & Crossley, H. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, *14*, 61–80.

Traugott, M., & Katosh, J. (1981). The consequences of validated and self reported voting measures. *Public Opinion Quarterly*, *45*, 519–535.

Reference Period

The reference period is the time frame for which survey respondents are asked to report activities or experiences of interest. Many surveys intend to measure frequencies of events or instances within a given period of time; for example, *How many times did you consult a medical practitioner during the last two months?* or *Think about the 2 weeks ending yesterday—have you cut down on any of the things you usually do about the house, at work, or in your free time because of illness or injury?* Most of the time, the reference period starts at some point in the past and ends at the time of the survey. However, there are fixed reference periods as well—for example, a calendar year or a calendar quarter, depending on the design of the study. Whereas some ongoing surveys (rotating panels or ongoing cross-sectional surveys) aim to document change throughout the field period, others intend to measure the incidence or prevalence of certain events within a given period of time.

In addition to the relative position of the reference period in relation to the time of the interview, its length is of key interest for survey researchers. The length—for example, number of days, weeks, or months—affects the variance of the estimated frequency (prevalence) or the proportion of respondents who have experienced a certain event (incidence). The longer the reference period is, the more stable is the estimate in reducing the variance associated to the variable. Thus, considering a given level of precision, survey cost is reduced. However, the length of the reference period depends on the concept to be measured; in case of a high prevalence variable (e.g., number of restaurant visits) a relatively short reference period might be appropriate, whereas the same reference period would not be appropriate for events with a relatively low frequency of occurrence (e.g., crime victimization).

Whereas a lengthy reference period seems preferable in terms of the variance of the estimate, it has considerable drawbacks.

1. Measurements of past events are subject to recall error. Recall error consists of various components of which *recall loss* and *telescoping* are the predominant ones. Recall loss is due to respondents forgetting certain events or instances that actually happened within the reference period, which in turn reduces the estimate compared to the true value. By contrast, (forward) telescoping produces higher estimates because respondents accidentally place an event into the reference period although it actually happened before the starting point of the reference period. To compensate for recall loss, researchers make use of introductory questions that stimulate memory by asking for autobiographical events within the reference period or other landmark events. By contrast, telescoping may be dealt with by bounded recall. To reduce telescoping, respondents are asked in two or more consecutive interviews to report on the frequency of certain events. At the time of the second measurement, respondents are asked to report on the

number of events or instances since the last interview, which serves as the outer "bound" for recall. Thus, events or instances reported in the first interview can be excluded from the response at the time of the second interview. This expensive procedure addresses the problem of respondents' tendency to "telescope" too distant experiences into the reference period.

2. Besides recall loss and telescoping, lengthy reference periods have additional disadvantages because they increase the time until results can be reported. For example, in an ongoing panel study on crime victimization, a 12-month reference period requires a longer period of time between the first occurrence of a given victimization and the availability of the report.

3. When it comes to answering a frequency question, respondents make use of at least two different strategies to generate a reasonable response: They either recall every single event that has occurred in the reference period and count the number of instances, or they estimate the number of instances using various estimation strategies (rate-based estimation, guessing)—depending on the characteristic of the event in question (similar vs. dissimilar, regular vs. irregular). Generally speaking, "recall and count" is considered to be advantageous compared to "estimating" in terms of the validity of the response. Even though the literature offers a great variety of findings, it is safe to assume that with longer reference periods, the proportion of respondents who estimate the number of events increases.

Considering these implications, survey researchers must balance the size of the variance (and cost) against the biases and disadvantages associated with a lengthy reference period.

Marek Fuchs

See also Bounding; Respondent-Related Error; Telescoping

Further Readings

Cannell, C., Fisher, G., & Bakker, T. (1965). *A study of the reporting of visits to doctors in the National Health Interview Survey*. Vital and Health Statistics, Series 2, No. 6. Hyattsville, MD: National Center for Health Statistics.

Mathiowetz, N. (2000). *The effect of length of recall on the quality of survey data*. Paper presented at the fourth international conference on Methodological Issues in Official Statistics, Stockholm, Sweden. Retrieved February 23, 2007, from http://www.scb.se/Grupp/Omscb/_Dokument/Mathiowetz.pdf

Neter, J., & Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, *59*(305), 18–55.

Refusal

In survey research, a refusal occurs when a request to participate in a survey is declined. In addition, some respondents who do participate can refuse to answer a particular question (sometimes called item nonresponse) or can refuse to finish the survey (sometimes called a partial or breakoff interview). The refusal rate is calculated as the proportion of contacted people who decline the survey request. Some researchers include partial interviews as refusals in the refusal rate calculation, others do not.

Refusals are important because they are a major type of nonresponse that can potentially introduce error in survey estimates. With refusal rates increasing both in the United States and worldwide, the reasons for refusals and how they are handled are important concerns for survey researchers.

In telephone surveys, the vast majority of refusals in surveys come shortly after the phone has been answered. As such, these nonresponders can be at the household level or at the respondent level. If the refusal occurs before the appropriate respondent within the household can be determined (either by random selection or other eligibility requirements), it is considered a household-level refusal. If the appropriate respondent has been determined and he or she is the person actually refusing the survey, it is considered a respondent-level refusal. A partial interview or breakoff is a respondent-level refusal.

For in-person and telephone surveys, interviewers may use a refusal report form to record any discernable details about the household or respondent, such as gender, age, and race. If the sample is from a client list or panel sampling frame, researchers may be able to estimate parameters related to nonresponse (e.g., demographics, past purchase behavior, or other list variables). These parameters from refusers can then be compared with the obtained sample to estimate the presence and impact of any potential nonresponse bias. Differences can be adjusted through survey weighting.

In self-administered surveys (e.g., mail survey, Internet survey), refusals make up some portion of those who do not return the questionnaire (nonresponders). For most self-administered studies, researchers know very little about why they did not participate because there is no interaction with an interviewer; for example, did they never receive the questionnaire or did they receive it but refuse? Depending on how the sample was obtained or constructed, researchers may or may not know which respondents did not participate or any other details about them. If parameters from nonresponders of self-administered surveys can be determined, again they can be used to determine potential nonresponse bias and used in survey weighting adjustments.

The reasons for refusal are varied. It is often difficult for researchers to ascertain the reasons for refusals because many do little more than hang up the phone ("immediate hang ups") or simply never respond in any way in the case of mail and Internet sampling. If there is an interaction with the interviewer, refusers usually do not communicate much aside from declining. Some may cite objections due to invasion of privacy, being reached at a bad time, length of interview, topic saliency, poor past survey experience, or a belief that the request is a telemarketing effort. Others may have a language barrier that prevents them from participating effectively. If the interviewer can determine (or at least suspect) a language barrier exists, this household can be recontacted by a bilingual interviewer.

Regardless of the reasons given, often the main reason for a refusal is that the interviewer has reached the household at an inconvenient time. In most mediated surveys, interviewers are initiating the contact with the household by proactively contacting it rather than a respondent returning an interviewer's call to take the interview, usually on a toll-free number. As such, the interviewer's request is often "interrupting" something at the household level. Unless specifically asked not to, researchers will typically recontact households that refused initially and will make a second request of the household to participate at a later date in the field period. Perhaps the interviewer will reach the household at a more convenient time or even talk to another household member than originally refused, possibly resulting in a completed interview. If a household refuses a second time, it is usually considered a "final" refusal and is not contacted again.

Although the Federal Trade Commission exempted survey and opinion research from the National Do Not Call Registry guidelines, most telephone survey organizations maintain internal do-not-call lists. Therefore, if a refuser asks to be placed on the do-not-call list, this information is recorded and this household will not be contacted again by that organization. Technically, if this household number is sampled in a subsequent survey by the same organization, it should be coded as a refusal even though it was not contacted.

In self-administered surveys, nonresponders are often sent reminders or additional requests to complete the survey. This could be a second copy of the questionnaire sent in the mail (the respondent could have misplaced the first one) or, in the case of an email invitation, a link to the online survey. Again, if respondents refuse and request not to be contacted further, they do not receive reminders or future survey requests. In fact, in most online surveys the invitation to participate includes simple instructions for how to opt out of any future requests or reminders.

Because handling refusers and potential refusers is a key part of an interviewer's job, training is especially important. A successful survey completion depends, in part, on the rapport established between the interviewer and the respondent. Poorly trained interviewers with nonneutral attitudes can hurt this relationship and lead to increased refusals. And given that the vast majority of refusals in telephone surveys occur in the first seconds of the interviewer–respondent interaction, the interviewer has very little time to develop rapport, anticipate potential barriers, and alleviate respondent objections. Refusal avoidance training focuses on how to avoid refusals by detecting respondent objections and proactively addressing them in an effort to persuade respondents to participate.

In most in-person surveys, interviewers have more time to develop rapport before the household member who opens the door refuses to cooperate. Furthermore, advance contact is more effective in in-person surveys as all sampled addresses can be mailed an advance letter. In telephone surveys of the U.S. general public only about half of the residences that are sampled can be matched to an accurate mailing address; but when this is possible sending an advance letter prior to placing the first call reduces the proportion of refusals appreciably.

In addition to learning how to disarm a potential refusal situation, interviewers can also be trained to "convert" initial refusals into completed interviews. This technique is called *refusal conversion*. Typically, sample households who have initially refused a survey request are called back by an experienced and trained interviewer. Some interviewers who are particularly adept at this type of respondent interaction may even be considered refusal conversion specialists. Converting refusals is important because it reduces nonresponse, saves costs, and may also reduce the potential bias it can introduce.

Sandra L. Bauman

See also Advance Contact; Advance Letter; Do-Not-Call Registries; Federal Trade Commission (FTC) Regulations; Household Refusal; Interviewer Training; Language Barrier; Missing Data; Nonresponse; Nonresponse Bias; Nonresponse Error; Nonresponse Rates; Partial Completion; Privacy; Refusal Avoidance; Refusal Avoidance Training (RAT); Refusal Conversation; Refusal Report Form (RRF); Respondent–Interviewer Rapport; Respondent Refusal; Response Rates

Further Readings

Brehm, J. (1993). *The phantom respondents: Opinion surveys and political representation*. Ann Arbor: University of Michigan Press.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Stec, J. A., & Lavrakas, P. J. (2007, May). *The cost of a refusal in large national RDD studies*. Paper presented at the 62nd annual conference of the American Association for Public Opinion Research, Anaheim, CA.

Refusal Avoidance

Because refusals make up a large portion of survey nonresponse, researchers want to minimize their occurrence as much as possible. Refusal avoidance is the researcher's awareness of, and efforts to eliminate or mitigate, factors that influence potential respondents toward refusing an invitation to participate in a survey. The relevance of concerted efforts to lower refusal rates is that the researcher is trying to reduce survey error through improvement in response rates, thus ostensibly improving the overall quality of the research.

Although there are many factors that influence a potential respondent's decision to participate in research, it is logical and cost-effective for researchers to concentrate on those factors under the control of the researcher, such as survey design and administration. As an integral part of the survey design process, the researcher not only should provide positive influences to a respondent to encourage participation (e.g., persuasive introductory text or verbiage, visually appealing research material, use of noncontingent monetary incentives) but also should strive to reduce or eliminate negative influences. Once the survey has been designed, the administration of the survey can also be refined with refusal avoidance in mind (e.g., interviewer selection, training, survey procedures, timing).

Seeking to increase the positive aspects during survey design is a natural starting point to avoid refusals and thereby improve response rates. Crafting introductory verbiage and responses (i.e., persuaders and other fallback statements) to respondents for telephone or in-person recruitment and the textual and visual appeal of mail or Internet recruitment will help avoid refusals and thereby improve the likelihood of response. This may include pretesting and focus groups with potential respondents or effective interviewers. Additionally, increasing the visibility or positioning of other positives should be considered, such leveraging (a) the research topic or sponsor, (b) that the entire population or certain subpopulations will benefit from the research results, (c) that incentives being are being offered, and (d) that by cooperating one is representing one's community.

Decreasing negative elements to research requests also is important, though it may not receive as much attention as increasing the positives. Often researchers will review their research materials and survey design to eliminate obvious negatives, such as confusing language. However, the researcher should also concentrate on understanding the population of interest, their social environment, and social-psychological attributes. For example, surveys among the various segments of the Asian population must be sensitive to their cultural heritage, such as knowing that many Chinese respondents may consider the number "4" (e.g., a $4 incentive) bad luck whereas 8 is considered a good number, white color is less favorable than red, and a gift of a clock (i.e., as an incentive) is a bad omen.

A survey should be administered with the aim of preventing refusals; both the delivery mode of the survey request(s) and the rules for making requests should be devised with this aim in mind. The researcher may improve likelihood of participation by attending to interviewer selection, training, and appearance; materials and delivery method (email, U.S. Postal Service, UPS, FedEx, etc.); visual appeal (use of color, pictures, logos, etc.); and so forth. The study contact rules can also improve response and avoid refusals through using effective contact times (e.g., evenings and weekends), number of contact attempts, using multiple modes for making the survey request, and so on. Additionally, the researcher should evaluate the reasons for refusals through use of a refusal report form or other means during and after the field period to improve future research (e.g., debrief meetings or focus groups with interviewers to catalog specific objections, areas of concern, and breakoff points during request or survey).

Charles D. Shuttles

See also Fallback Statements; Interviewer Training; Nonresponse; Refusal; Refusal Avoidance Training (RAT); Refusal Rate; Refusal Report Form (RRF)

Further Readings

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2002). *Survey nonresponse*. New York: Wiley.

Groves, R. M., & McGonagle, K. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17*(2), 249–266.

Groves, R. M., Singer, E., & Corning, A. C. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, *64*, 299–308.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Refusal Avoidance Training (RAT)

Research interviewers have the difficult task of obtaining the cooperation of respondents. Successful interviewers are able to achieve positive outcomes (completed interviews) while simultaneously avoiding negative outcomes (refusals). Researchers may employ several approaches to improving response rates through refusal avoidance, including the use of refusal avoidance training (RAT), which specifically concentrates on interviewers reducing the proportion of their survey requests that end as a refusal.

Experienced and successful interviewers tailor their approach to individual respondents, rather than using a "one size fits all" approach. To successfully tailor their approach, interviewers must have an extensive set of techniques, strategies, phrases, and so on, that have to be customized to the specific survey request. Interviewers must use *active listening skills* to pick up on the verbal and nonverbal cues of the respondent. These cues will assist the interviewer in selecting the appropriate response strategy most likely to elicit respondent cooperation.

Interviewers skilled at maintaining interaction (continuing their contact with the respondent) create more opportunity to tailor their approach. As the interaction between the interviewer and the respondent continues, and the time investment grows longer, it becomes more difficult, in theory, for the respondent to break off contact and refuse the survey request.

An innovative approach to the development and implementation of refusal avoidance training was first posited by Robert Groves and K. McGonagle. A three-step process is used to develop a customized (to the survey organization, topic, and sponsor) refusal avoidance training program. The steps are summarized as follows:

1. Focus groups of experienced interviewers are held to capture specific examples of the actual words used by reluctant respondents to describe their concerns about the survey request. Focus group moderators seek to maximize the number of different types of concerns recalled by the interviewers. Hundreds of utterances from respondents may be collected.

2. After assembly and elimination of duplicate concerns, senior interviewers and training staff classify the concerns into thematic sets (e.g., concerns about privacy, insufficient time), and then identify the desirable verbal behaviors of interviewers to address the concerns. There are often multiple alternative behaviors that may be used by the expert interviewers in response to a specific utterance; each, however, addresses the expressed concern(s) of the respondent.

3. A training curriculum can then be developed and concentrated on, imparting four refusal avoidance skills:
   a. Learning the themes of a potential respondent's concerns.
   b. Learning to classify a potential respondent's actual wording of a concern into those themes (the diagnosis step).
   c. Learning desirable verbal behaviors to address the concerns.
   d. Learning to deliver to the reluctant respondent, in words compatible with their own, a set of statements relevant to their concerns.

A major goal of the refusal avoidance training is to increase the speed of the interviewer's performance on points 3b through 3d.

Additional refusal avoidance skills go beyond quickly and effectively responding to respondent reluctance by concentrating on the opening 5 to 10 seconds of contact with the respondent. An effective introduction, and training to improve the introduction, focuses on interviewer alertness, prior knowledge, perceiving the nature of the respondent's "hello," and active listening. Effective interviewers are noted for being on high alert for their next survey request. Interviewers should specifically prepare themselves to react to the possible outcomes and reactions of requesting survey participation. Part of being prepared is to glean as much information as possible prior to the request, such as knowledge of the respondent's location (state, city, neighborhood, etc.), knowledge of prior attempts, and the like. Once the interviewer initiates a survey request, there are the verbal cues (and visual cues in the case of face-to-face interviewing) that result from the potential respondent's greeting (i.e., the hello). Does the respondent appear to be impatient? tired? curious? hostile? Interviewers should employ active listening for additional background cues to further tailor their introduction; for example, sounds of activity (kids, household guests, loud music, or television) could prompt the interviewer to use a briefer introduction.

Researchers can improve the effectiveness of their training techniques by incorporating large group, small group, paired, and individual exercises and role playing activities that concentrate on increasing interviewers' response speed and appropriate selection of effective responses to reluctant respondents. If effectively developed and administered, refusal avoidance training may substantially increase an interviewers' "toolbox" of possible effective responses to stated concerns from reluctant respondents and ingrain quick responses.

Charles D. Shuttles

See also Interviewer Training; Refusal; Refusal Avoidance; Tailoring

Further Readings

Groves, R. M., & McGonagle, K. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17*(2), 249–266.

Shuttles, C. D., Welch, J., Hoover, B., & Lavrakas, P. J. (2002, May). *The development and experimental testing of an innovative approach to training telephone interviewers to avoid refusals*. Paper presented at the 59th annual conference of the American Association for Public Opinion Research Conference, St. Petersburg, FL.

Refusal Conversion

Refusal conversions are the procedures that survey researchers use to gain cooperation from a sampled respondent who has refused an initial survey request. Refusal conversion may include different versions of the survey introductions and other written scripts or materials (e.g., cover letters), study contact rules, incentives, and interviewer characteristics and training. This is a common procedure for many surveys, but it requires careful consideration of the details of the refusal conversion efforts and the potential costs versus the potential benefits of the effort.

The goal of converting initial refusals is to raise the survey response rate, under the assumption that this may lower the potential for refusal-related unit nonresponse error. The research literature contains reports of successfully converting refusals in telephone surveys between 5% and 40% of the time. There is little reported in the research literature about whether refusal conversions efforts are effective in reducing nonresponse bias.

Gaining cooperation from a potential respondent during initial contacts is much more effective than attempts to "convert" that respondent after an initial refusal has been encountered. Thus, all researchers should pay attention to survey design and administration to effectively maximize cooperation and minimize refusals for initial contacts. However, despite researchers' best efforts to avoid refusals, they will still occur; thus, refusal conversion is an avenue to gain some amount of completed questionnaires from initial refusers. The basic procedures used to carry out the administration of the survey may need to be modified during refusal conversion attempts.

Refusal conversions often are attempted in surveys that are conducted via mail or the Internet. In these survey modes, refusal conversion essentially is part and parcel of the larger process of recontacting respondents who have not yet returned a completed questionnaire under the assumption that some portion of them initially have decided to refuse to do so. Procedures for doing these follow-up contacts of initial nonresponders in mail and Internet surveys are described elsewhere in this encyclopedia and will not be discussed here in any detail. Instead, the remainder of this entry focuses mostly on refusal conversion as applied in face-to-face and telephone interviewing.

Approaches

There are two basic ways to structure the refusal conversion process: (1) ignoring the fact that a respondent has refused an initial survey request and using the exact same survey request approach, or (2) implementing a revised or different approach when trying to convert the initial refusal. The first approach relies upon the hope that the timing will be more opportune or the respondent's inclination to participate will be more favorable when recontacted than at the time of the first contact. When using this approach, the researcher should modify survey procedures by designing recontact rules that use as much lag time (i.e., amount of time between the initial refusal attempt and the subsequent refusal conversion attempt) as possible. If the field period allows, lag time of 5 to 14 days has been found to increase the likelihood of encountering a different respondent (in household or organizational surveys) or reaching the original respondent at a more convenient or favorable time.

Many survey organizations use the second approach of doing a more extensive modification of survey procedures for subsequent contacts to a respondent after the initial refusal. For mail or Internet mode written survey requests, this may mean using different text in cover letters that acknowledges the previous attempt, addresses the possible reason for refusing (e.g., privacy, survey sponsor, timing, legitimatizing research purpose[s]), and explaining the importance of participation (e.g., how a government program may benefit from higher research participation). For in-person or telephone surveys, this may mean using different introductory scripts that vary the appeal. Additionally, changing incentives, (e.g., increasing the incentives if lower amounts were initially offered) or offering different types of incentives (i.e., contingent vs. noncontingent) may be effective tactics to use during refusal conversion attempts. Another approach for researchers to consider is to offer a shortened questionnaire (containing only key questions) to reduce the burden on the respondent.

The contact or calling rules also can be varied to improve the likelihood of success during refusal conversion attempts, including mode of contact, interviewer assignment, advanced interviewer training, refusal conversion specialists, and gathering special information from the interviewer at the time of initial refusal, for example, through the use of a refusal report form.

Mode of Contact

If an initial refusal occurs in one mode (e.g., a mailed survey), the use of one or more other modes can be considered if cost and time permit. For example, a refusal to an initial mailed survey request could spur in-person or telephone refusal conversion attempts.

Interviewer Assignment

The attributes and skills of the original in-person or telephone interviewer may be related to why he or she did not gain the cooperation of the respondent during the initial survey request. By assigning a different "type" of interviewer in subsequent refusal conversion attempts, cooperation may be gained through different skills or a different approach of that interviewer. Some researchers believe it is wise to match demographic characteristics as closely as possible when deciding which interviewers will try to recontact an initial refuser. This, of course, can be done only if demographic information is available about the refuser, such as what can be captured in a refusal report form.

Refusal Report Form

The refusal report form contains information that can be collected at the time of the initial refusal in face-to-face and telephone surveys and can be extremely beneficial in assisting the subsequent refusal conversion attempt. Specifically, a refusal report form can be used to collect information such as (a) perceived respondent demographics (e.g., sex and age), (b) degree of interaction, (c) reasons and concerns given for refusing, (d) the strength of the refusal, (e) whether the refusal was at the household or respondent level, (f) when the refusal took place (e.g., immediate refusal, middle of introduction, after introduction, during data collection), and (g) other visual and audio cues (e.g., condition of household, neighborhood characteristics, presence of children). In this way, the interviewer performing the refusal conversion attempt can better prepare in advance of recontacting the initial refuser by using this information to tailor the approach during the conversion attempt to that particular refuser and his or her initial reasons for refusing.

Refusal Conversion Training

This type of advanced training concentrates on refusal conversion. The development of this training may be similar to, or combined with, refusal avoidance training.

Refusal Conversion Specialists

Evidence shows that some interviewers are better at refusal conversion than are others. Granted, researchers should concentrate on having interviewers who are all skilled at effectively gaining cooperation on the initial contact and thereby avoid refusals as often as possible. This notwithstanding, proper selection of effective refusal conversion specialists can maximize refusal conversions over what can be achieved by the general pool of interviewers. Attributes of successful refusal conversion specialists are (a) confidence, (b) perseverance in dealing with reluctant respondents, (c) appropriate assertiveness, (d) skill at tailoring and using multiple approaches, and (e) quick adaptation and response to whatever the reluctant respondent is saying. Refusal conversion can be assigned to supervisors or other interviewing specialists that often receive higher remuneration or incentives for taking on these especially difficult cases.

Recontact Determination

Deciding which initially refusing cases to try to convert is also an important part of the refusal conversation process. In almost all interviewer-administered surveys, the interviewer should never attempt to convert all initial refusers, because some of them will have refused so vehemently or pointedly as to make recontact wholly inappropriate. For example, it is generally agreed among survey professionals that any initial refuser who states something to the effect, "Don't ever contact me again!" should not be subjected to a refusal conversion attempt. The use of a refusal report form can be very helpful in determining which initial refusers to try to convert. This can be done manually by supervisory personnel reviewing the prior refusals, or if the data from the refusal report form are entered into a computer, an algorithm can be devised to select those initial refusers who should be tried again and select out those that should not be contacted again.

Costs Versus Benefits

Finally, the researcher should carefully consider the potential costs versus the benefits of trying to convert refusals. Refusal conversion efforts can span a wide spectrum of effort and contacts with a varied return on investment. For example, a possible outcome of too aggressive a refusal conversion process may be a decline in overall data quality. Thus, those initially refusing respondents who are extremely reluctant but who agree to be interviewed during the conversion contact may not be interested in the topic, the research sponsor, or in contributing to the outcome of the research and thus may provide low-quality data. By employing refusal conversion procedures, the researcher may motivate respondents just enough to participate but with increased item nonresponse or other types of low-quality responses (e.g., converted refusers may disproportionately employ satisficing, by giving answers that seem reasonable but without searching their memory or giving adequate cognitive consideration).

There has been little research on the costs of refusal conversions. But in 2007, Jeffrey A. Stec and Paul J. Lavrakas reported findings based on several very large random-digit dialing surveys that suggested that completed interviews gained from converted refusals are far less costly to obtain than a completed interview from a yet-to-be-attempted telephone number or address. If this finding holds up in other studies, it would indicate that refusal conversions not only increase survey response rates but also lead to cost savings compared to adding new samples to compensate for all the cases that were lost due to initial refusals.

Charles D. Shuttles, Paul J. Lavrakas, and Jennie W. Lai

See also Calling Rules; Cover Letter; Nonresponse Bias; Nonresponse Error; Refusal; Refusal Avoidance; Refusal Avoidance Training; Refusal Report Form (RRF); Respondent Burden; Unit Nonresponse

Further Readings

Bauman, S. L., Lavrakas, P. J., & Merkle, D. M. (1994, May). *Data quality of reluctant respondents: Refusal conversions and item-nonresponse*. Paper presented at 49th annual conference of the American Association for Public Opinion Research, Danvers, MA.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.). (2002). *Survey nonresponse*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Lavrakas, P. J., Bauman, S. L., & Merkle, D. M. (1992, May). *Refusal report forms, refusal conversions, and nonresponse bias*. Paper presented at the 47th annual conference of the American Association for Public Opinion Research, St. Petersburg, FL.

Stec, J. A., & Lavrakas, P. J. (2007, May). *The cost of a refusal in large national RDD studies*. Paper presented at the 62nd annual conference of the American Association for Public Opinion Research, Anaheim, CA.

Refusal Rate

The refusal rate is the proportion of all potentially eligible sample cases that declined the request to be interviewed. Before calculating the refusal rate, researchers must make some decisions about how to handle the types of nonresponse in the calculation.

There are two types of refusals that are generally included in the refusal rate calculation:

1. *Household-level refusals:* These refusals occur before the appropriate respondent within the household can be determined (either by random selection or other eligibility requirements). That is, the person refusing may or may not be the selected respondent within that household.
2. *Respondent-level refusals:* These are refusals that occur after the appropriate respondent has been determined. That is, the selected respondent has been identified and that individual is the one actually refusing.

Partial interviews ("partials"), or breakoffs, can be classified as another type of respondent-level refusal. Partials are when respondents begin an interview but do not finish it. They may be refusing to complete the questionnaire, or they may have other circumstances that interfere with finishing it (e.g., they initially ran out of time and were not recontacted to finish the interview during the field period). Depending on how many questions were answered, researchers may treat some breakoffs as completed interviews and others as respondent refusals or noncontacts as a final disposition.

There are three generally accepted refusal rate calculations that are included in the American Association for Public Opinion Research's 2006 Standard Definitions. In all three refusal rate calculations, the numerator is simply the total number of refusals (household- and respondent-level). But each of the calculations differs according to which dispositions or call outcomes are included in the denominator.

Refusal Rate 1. The denominator includes all possibly eligible cases still in the sample frame, regardless of whether or not the eligibility of the case could be determined. Thus, the denominator includes completed interviews, refusals, noncontacts (e.g., callbacks), and cases of unknown eligibility. This rate is the most conservative of the three refusal rate calculations.

Refusal Rate 2. The denominator is similar to Refusal Rate 1 except it uses only a proportion of the unknown cases, not all of them. This proportion is an estimate (e) of how many unknown cases would likely be eligible (e.g., is a household and meets other survey criteria). In estimating this proportion, researchers must be guided by the best available information on what share of eligible cases is likely among the unknown eligibility cases.

Refusal Rate 3. The denominator excludes all cases of unknown eligibility. In other words, this rate is the proportion of refusals among those sample cases known to be eligible sample units for the survey. Noncontact and other rates can be calculated in a manner similar to this rate so that, when summed, all will equal the total nonresponse rate.

The survey refusal rate has been increasing in the past 20 years in the United States and elsewhere in the world. This has happened despite the efforts of many survey researchers around the world. There is no easy or apparent solution to this decline other than to continue to investigate its causes and possible solutions.

Sandra L. Bauman

See also *e*; Household Refusal; Noncontacts; Nonresponse Bias; Partial Completion; Refusal; Respondent Refusal; Standard Definitions; Unknown Eligibility; Within-Unit Selection

Further Readings

The American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Refusal Report Form (RRF)

A large proportion of nonresponse in surveys is due to refusals, which occur when a request to participate in a survey is declined. Researchers are concerned about the effect of refusals because of the potential error that they can introduce in survey estimates.

A refusal report form (RRF) is a structured form used by interviewers immediately after a refusal is encountered. The employment of an RRF process produces valuable paradata and has three main benefits:

1. It can provide estimates of relevant parameters (e.g., refuser gender, race, age) that can be compared with the obtained sample to help determine the presence and impact of potential nonresponse bias.
2. It can provide valuable information that can help interviewers in subsequent contact attempts to convert these refusals into completed interviews.
3. It can help researchers conduct investigations into the nature of refusals so as to plan better strategies on how to reduce their frequency of occurrence.

RRFs capture structured information about all individual refusals that most often is lost when interviewers are given only informal instructions to write down notes about the refusals if they think it is appropriate.

RRFs are used in mediated interviews—that is, those that are conducted by interviewers, either in person or on the phone. Information about the refusal is recorded based on estimates that interviewers make either visually (for in-person) or audibly (for telephone and in-person).

There is no standardized format for an RRF. (Paul J. Lavrakas appears to have been the first to describe the use and value of such a form and to show an example of it in a book on telephone survey methods.) Researchers develop RRFs with variables that are most relevant to the study and that are reasonable to be estimated by interviewers given the situation. For example, in an in-person survey, an interviewer may be able to provide information about the type of home, neighborhood setting, and so forth. Telephone interviewers could not begin to estimate those details from brief phone conversation but past research has shown that they can provide accurate estimates of certain demographic characteristics for the person being spoken to.

In any case, the RRF usually tries to capture two types of information, linked to the two benefits specified previously. Demographic information about the refuser (e.g., gender, age, race) and details about the context of the refusal (e.g., strength of the refusal, reasons given for refusal, perceived barriers to participating, etc.) may help interviewers, in future attempts, to convert that refuser into a cooperating respondent.

Given that the vast majority of refusals in telephone surveys typically occur in the first few seconds of the interviewer–respondent interaction, the interviewer has very little time to develop rapport, anticipate potential barriers, and alleviate respondent objections. Although there has been little published about the use of RRFs, research by Sandra L. Bauman, Daniel M. Merkle, and Paul J. Lavrakas suggests that telephone interviewers can accurately make estimates of gender, race, and age in a majority of cases, even when the interactions are brief. These estimates can be used to help determine the presence of nonresponse bias.

Interviewers are also able to provide details about the refusal that can help survey management determine which refusals to contact again and how best to approach that next interaction. For example, if an interviewer is recontacting a refusing household where the reason given was "We're just sitting down to dinner," they may start the next interaction with "I'm sorry we reached you at a bad time. Is this better?" Or, if the RRF indicates that a woman refused but during the conversion attempt a man answers, the interviewer can adjust the introduction accordingly.

When potential respondents refuse to cooperate, researchers know little about them and why they choose to not participate for the simple reason that they refuse to talk to the interviewer. In in-person surveys, more information about the refuser is at the interviewer's disposal because he or she actually saw the person who refused. The problem of gathering information on refusers is an especially challenging one for telephone survey researchers; often the best proxy is interviewer estimates (like those gathered via RRFs).

Sandra L. Bauman

See also Nonresponse Bias; Paradata; Refusal; Refusal Conversion; Refusal Rate; Respondent–Interviewer Rapport

Further Readings

Bauman, S. L., Merkle, D. M., & Lavrakas, P. J. (1992, November). *Interviewer estimates of refusers' gender, age, and race in telephone surveys*. Paper presented at the 16th annual conference of the Midwest Association for Public Opinion Research, Chicago, IL.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Lavrakas, P. J., Bauman, S. L., & Merkle, D. M. (1992, May). *Refusal report forms, refusal conversions, and nonresponse bias*. Paper presented at the 47th annual conference of the American Association for Public Opinion Research, St. Petersburg, FL.

Registration-Based Sampling (RBS)

Registration-based sampling (RBS) is a sampling frame and sampling technique that has been used, with growing frequency in the past two decades, for conducting election polls. RBS frames for a given geopolitical area can be built by researchers using public records for that political jurisdiction, or they can be purchased from vendors who already have done the legwork. Unlike the random-digit dialing (RDD) telephone sampling frame that has been used primary for election polling since the 1980s, the RBS frame is comprised of a list of names, addresses, and oftentimes telephone numbers of registered voters. An immediate advantage of RBS is that the name of the sampled voter is available for use in gaining that respondent's cooperation. Another major advantage of the RBS frame over the RDD frame is that RBS often comes with other valuable variables to help plan the sampling design that will be used for an election poll. These variables include information about the voter such as age, political party affiliation, and past voting frequency. A major disadvantage of the RBS frame compared to RDD is that the quality of RBS varies considerably across different jurisdictions, and coverage of the probable electorate can be so poor as to render the RBS frame invalid in some jurisdictions.

A major challenge faced by those who conduct polls to predict (forecast) an election outcome, and by those who study voters after an election, is to accurately identify who will vote or who has voted. Pre-election pollsters have created many approaches for use with RDD sampling to screen their samples for so-called likely voters who will make up the probable electorate. These approaches are imperfect and often do not work well, thereby contributing to inaccuracies in election outcome predictions. With RBS that uses an enhanced database that includes a registered voter's past voting frequency and party affiliation, a model can be devised not only to better predict the likelihood someone actually will vote but also to better predict for which candidate the person will vote, in the case of those who have declared a party affiliation. With such information appended to the RBS frame about each registered voter, an RBS researcher also can stratify the sample and make more cost-effective decisions about how many voters to interview who have declared a party affiliation versus those who have not (i.e., the independents).

When sampling from an RBS frame, the researcher will generally segment the frame into three strata: (1) those who voted in the past election(s), (2) those who were registered but did not vote in the past election(s), and (3) those who were not registered for the past election(s). Based on a number of auxiliary sources of information, the researcher then will estimate the proportion of registered voters in each of these groups who are expected to vote. Using all this information, the researcher will make decisions about how many voters to sample from each strata.

Because RBS frames have addresses, and often telephone numbers, for each registered voter, the mode of data collection can be mail, telephone, in-person, or any combination of these. Not enough methodological work has been done with RBS to conclude with confidence under what circumstances it should be used as opposed to RDD sampling, but as public records become more uniform in their quality, the field of election polling can expect to see an increasing use in RBS and a decreased use of RDD.

Paul J. Lavrakas

See also Coverage Error; Election Polls; Likely Voter; Probable Electorate; Random-Digit Dialing (RDD); Sampling Frame

Further Readings

Green, D. P., & Gerber, A. S. (2003, May). *Using registration-based sampling to improve pre-election polling*. Paper presented at the 58th annual conference of the American Association for Public Opinion Research, Nashville, TN. Retrieved April 10, 2008, from http://bbs.vcsnet.com/pdf/registrationbasedsampling_Smith-Richardson-Report_11_10_03.pdf

Regression Analysis

Regression analysis is the blanket name for a family of data analysis techniques that examine relationships between variables. The techniques allow survey researchers to answer questions about associations between different variables of interest. For example, how much do political party identification and Internet usage affect the likelihood of voting for a particular candidate? Or how much do education-related variables (e.g., grade point average, intrinsic motivation, classes taken, and school quality) and demographic variables (e.g., age, gender, race, and family income) affect standardized test performance? Regression allows surveyors to simultaneously look at the influence of several independent variables on a dependent variable. In other words, instead of having to calculate separate tables or tests to determine the effect of demographic and educational variables on test scores, researchers can examine all of their effects in one comprehensive analysis.

Regression also allows researchers to statistically "control" for the effects of other variables and eliminate spurious relationships. In a more serious case, a case of noncausal covariation, two variables may be highly related but may not have a direct causal relationship. For example, in cities in the United States, murder rates are highly correlated with ice cream sales. This does not mean, however, that if the selling of ice cream is curtailed that the murder rate will go down. Both ice cream sales and murder rates are related to temperature. When it gets hot out, people buy more ice cream and commit more murders. In a regression equation, both ice cream sales and temperature can be included as predictors of murder rates, and the results would show that when temperature is controlled for, there is no relationship between ice cream sales and murder rates.

This ability to control for other variables makes arguments based on research results much stronger. For example, imagine that a test score regression showed that the more English classes a school required, the better their students did on standardized tests, controlling for median family income, school quality, and other important variables. Policy advocates can then propose increasing the required English courses without being as open to the criticism that the results were really due to other causes (such as socioeconomic status).

The regression approach can also simultaneously look at the influence of different important variables. For example, imagine that the head reference librarian and the head of acquisitions for a library disagree about whether it is customer service or having the most up-to-date bestsellers that influences patron satisfaction. A regression predicting patron satisfaction from both customer service ratings and percentage of recent bestsellers can answer the question of which one (or both or neither) of these factors influences customer service. Researchers can even look at interactions between the variables. In other words, they can determine if the effect of customer service on patron satisfaction is bigger or smaller at libraries with fewer bestsellers than those with more bestsellers.

At its base, the linear regression approach attempts to estimate the following equation:

$$y = b_1x_1 + b_2x_2 + \cdots b_nx_n + e,$$

where y is the dependent variable; $x_1, x_2 \ldots x_n$ are the independent variables; e is the error in prediction; and $b_1, b_2 \ldots b_n$ are the regression coefficients. The regression coefficients are estimated in the model by finding the regression lines that simultaneously best minimize the squared errors of prediction (i.e., the sums of squares). If a dependent variable, controlling for the effects of the other dependent variables, has a large enough relationship with the independent variable, then the regression coefficient will be significantly different from zero. Regression coefficients can be interpreted as partial slopes; in other words, the regression coefficient indicates that for each one-unit increase in the independent variable (and controlling for the effects of the other independent variables), the dependent variable increases or decreases by the amount of the regression coefficient.

Assumptions of Linear Regression

For a regression analysis to be valid, there are several assumptions that need to be satisfied. First, the errors must be independent and normally distributed. Nonindependent error terms often occur when there are relations between responses, such as responses from married couples or individuals from one household.

Somewhat obviously, it is important that the relationship between the independent variables and dependent variable are linear. Somewhat less obviously, it is important the errors of each independent variable have essentially the same variance (i.e., they do not exhibit heteroscedasticity).

Fortunately, regression is fairly robust to small violations of all of these assumptions. However, to be sure that a regression model is not affected by violating any of these assumptions, and therefore providing biased answers, researchers should ensure that they are using the correct technique and availing themselves of the regression diagnostic measures (primary among which include plotting both the data and the residuals) that any standard statistical software package provides.

Other Types of Regression

Many types of regression beyond linear regression have been developed to deal with special cases of analysis or for situations where using linear regression would result in a gross violation of its assumptions. While the description of all of the varieties of regression is beyond the scope of this entry, several common and useful methods are mentioned briefly.

Logistic regression is designed to handle categorical dependent variables (i.e., Yes–No questions or other cases without a continuum of answer options). Because categorical dependent variables do not have error distributions, logistic regression uses the logit function to transform the analysis into an examination of how independent variables affect the odds of the occurrence of a particular dependent variable response option choice. Because the results are expressed in odds and odds ratios, they can often be challenging to interpret.

Hierarchical linear models are a family of techniques designed to deal with data that have nonindependent errors. For example, if students from different classrooms from one school were sampled for a survey on preferred pedagogical techniques, responses from students from the same classroom would have similar answers because of their exposure to the same teacher. A hierarchical linear model would create a model where students (within classrooms) was the unit of analysis and another where classroom was the unit of analysis, looking at appropriate independent variables at each level.

Limits of Regression Analyses

Like any form of research, regression is not a panacea. In addition to the assumptions discussed previously, there are several limitations that researchers should be aware of, including (among others) the danger of the ecological fallacy, the issue of mis-specifying a model or not including all of the relevant predictors, or being led astray by an idiosyncratic sample. The foremost limitation is understanding that there is a difference between a relationship and a causal relationship. Even if there are statistically significant relationships between independent and dependent variables, that does not mean that there is necessarily a direct causal relationship between the variables. It is possible that there is a relationship but that the chain of causality is rather long or indirect. In the ice cream and murder example discussed earlier, high temperatures do not directly cause more murders. Higher temperatures likely cause more discomfort, which probably causes more frustration, which likely leads to lower thresholds for anger, which then probably leads to more violence, which leads to more murders. However, even in the absence of direct causality, regressions are a powerful tool that, with the selection of the proper control variables, can shed light on important relationships between variables and, with good statistical confidence, can examine the effects of one variable on another.

Geoffrey R. Urland and Kevin B. Raines

See also Alpha, Significance Level of Test; Analysis of Variance (ANOVA); Confidence Interval; Correlation; Dependent Variable; Ecological Fallacy; Independent Variable; Interaction Effect; Mean Square Error; Noncausal Covariation; Outliers; *p*-Value; Statistical Power; *t*-Test; Type I Error; Type II Error

Further Readings

Freedman, D. A. (1999). From association to causation: Some remarks on the history of statistics. *Statistical Science*, *14*, 243–258.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. San Diego, CA: Harcourt Brace.

REINTERVIEW

A reinterview occurs when an original respondent is recontacted by someone from a survey organization—usually not the original interviewer—and some or all of the original questions are asked again. Reinterviewing can serve more than one purpose in survey research, including (a) verifying that interviews were actually completed as the researchers intended with sampled respondents, (b) checking on the reliability of the data that respondents provided when they were originally interviewed, and (c) further studying the variance of survey responses.

As part of quality assurance efforts to monitor the quality of survey data collection when the questionnaire is interviewer-administered, some part of the recontacts made to selected respondents may include asking some of the original questions again, especially the questions that the researchers deem as key ones. Although there are many reasons why a respondent may not provide the exact same answer during the reinterview, including some legitimate reasons, the purpose of re-asking some of the questions is not to match answers exactly but rather to make sure there is no consistent pattern of deviation from the original data that could signal that the questionnaire was not administered properly by the interviewer, including the possibility that the interviewer falsified some or all of the data.

A reinterview also provides data that can be used by researchers to test the reliability of the original data. For example, demographic characteristics are unlikely to change if the reinterview is conducted within a few weeks or even a few months of the original data collection, although some might, such as someone turning one year older in age, or someone becoming a college graduate because of a recent graduation, or now becoming employed in a new occupation in the interim since first being interviewed. Other types of questions, such those concerning behaviors, experiences, perceptions, knowledge, attitudes, and opinions, also are unlikely to change much within a few weeks or even a few months, but they can be expected to be more likely to change than will demographics. Small changes in these types of variables do not necessarily mean the original data are unreliable, but large changes often signal problems with the quality of (1) the original interviewing, (2) the questionnaire wording, (3) the data collection performed during the reinterview, or all three of these factors. Data gathered from re-asking the same questions in the reinterview also provide researchers with additional ways to understand the variance that is associated with their questionnaire items.

Granted, reinterviewing is costly, but when survey budgets allow for it, there are many benefits that can be gained.

Paul J. Lavrakas

See also Falsification; Quality Control; Recontact; Reliability; Variance Estimation; Verification

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

O'Muircheartaigh, C. (1991). Simple response variance: Estimation and Determinants. In P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 551–574). New York: Wiley.

Relative Frequency

Relative frequency refers to the percentage or proportion of times that a given value occurs within a set of numbers, such as in the data recorded for a variable in a survey data set. In the following example of a distribution of 10 values—1, 2, 2, 3, 5, 5, 7, 8, 8, 8—while the absolute frequency of the value, 8, is 3, the relative frequency is 30% as the value, 8, makes up 3 of the 10 values. In this example, if the source of the data has a wider range of possible scores than the observed values (such as a 0–10 survey scale), then it is permissible to report that some of possible values (e.g., 0, 4, 6, 9, and 10) were not observed in this set of data and that their respective relative frequencies were zero (i.e., 0%).

In survey research, the relative frequency is a much more meaningful number than is the absolute frequency. For example, in a news article using results from a poll of 800 citizens, it is more meaningful to know that approximately two thirds of them (67.5%) are dissatisfied with the job the president is doing than to know that 540 citizens who were polled think this way.

Relative frequency can be displayed in a frequency table—which displays each value in a distribution ordered from lowest to highest—along with the absolute and cumulative frequencies associated with each value. Relative frequency also can be displayed graphically in a bar graph (histogram) or pie chart.

Paul J. Lavrakas

See also Frequency Distribution; Percentage Frequency Distribution

Reliability

The word *reliability* has at least four different meanings. The first of these is in an engineering context, where reliability refers to the likelihood that a piece of equipment will not break down within some specified period of time. The second meaning is synonymous with *dependability*, as in "She is a very reliable employee" (i.e., she is a good worker). The third has to do with the sampling variability of a statistic. A percentage, for example, is said to be reliable if it does not vary by any nonnegligible amount from one sample to another of the same size and that are drawn in the same manner from the same population.

The fourth meaning is the one focused upon in this entry. A measuring instrument is said to be reliable if it yields consistent results, whether or not those results are valid (i.e., whether or not the results are relevant to the purpose for which the instrument is intended).

The social science literature is replete with discussions of different kinds of measurement reliability. There is test–retest consistency (agreement from time to time using the same instrument); parallel forms consistency (agreement between one instrument and another interchangeable instrument); internal consistency (agreement among the items within an instrument); interobserver consistency (or *intercoder reliability*; agreement between one rater and another); and intra-observer consistency (agreement within the same rater from one occasion to another).

The key concepts in the classical theory of reliability are "observed score" (the measurement actually obtained), "true score" (the measurement that, in some sense, should have been obtained), and "error score" (the difference between true score and observed score); the latter two are generally unknown. The reliability coefficient for an instrument in a given study is defined as the ratio of the variance of the true scores to the variance of the observed scores, and it is estimated by (a) the agreement between observed scores at Time 1 and observed scores at Time 2; (b) the agreement between observed scores on Form A and observed scores on Form B; (c) the agreement among the observed scores on the items that constitute the instrument; (d) the agreement between the ratings given by Judge A and the ratings given by Judge B; or (e) the agreement between the ratings given by Judge A on Occasion 1 and the ratings given by Judge A on Occasion 2.

Two Hypothetical Examples in Survey Research

1. In a pilot study preparatory to the main study in which a self-report questionnaire of cigarette smoking behavior (concerning the number of cigarettes smoked per day) is to be employed, a survey researcher might administer a trial version of the questionnaire to a sample of respondents at two different times, with perhaps a few days in between, and determine the extent to

which the responses given by the participants at Time 1 agree with (i.e., are reliable compared to) the responses given by those *same* participants at Time 2.

2. In an effort to study the reliability of people's self-reports of their ages, a survey researcher might send an initial mailing of a postcard to a sample of persons whose names and addresses have been randomly sampled from a telephone directory, asking them to record their birth date on the stamped return-addressed portion of the postcard and mail it back. At some subsequent point in time (perhaps a month or so later), the researcher might field a telephone survey of the same people including a question asking them to report their date of birth. The second set of birth dates could be compared with the first set, on a respondent-by-respondent basis, to see how well the two agree.

Relative Versus Absolute Agreement

Reliability is usually assessed by correlating the scores at Time 1 with the scores at Time 2, correlating the scores on Form A with the scores on Form B, and so forth. But correlation is concerned only with the relative relationship between two variables. Scores of 1, 2, 3, 4, and 5 correlate perfectly with scores of 1, 3, 5, 7, and 9, respectively, but there is perfect agreement only for the first pair (1,1). If a measure of absolute agreement is desired, the researcher should use something like the median absolute difference between paired scores.

Reliability of an Instrument Versus Reliability of Scores Obtained With an Instrument

It is somewhat controversial whether one should refer to the reliability of an instrument or the reliability of scores obtained with the instrument. (Scores need not be test scores as such; they could be heights, weights, temperature readings, etc.) If it is clear from the context what the reliability information is for a given instrument in a given study, no great harm is done by referring to the reliability of the instrument itself.

Other Approaches to Reliability in the Social Sciences

The preceding discussion has been concerned with so-called classical reliability. In recent years there has appeared a variety of other approaches to the reliability of measuring instruments. One of these is based upon generalizability theory; another is based upon item-response theory; a third is based upon structural equation modeling.

Thomas R. Knapp

See also Intercoder Reliability; Item Response Theory; Test–Retest Reliability; Validity

Further Readings

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *15*(4), 21–33.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice*, *16*(4), 14–20.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Litwin, M. S. (1995). *The Survey Kit: Vol. 7. How to measure survey reliability and validity*. Thousand Oaks, CA: Sage.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8–14.

Repeated Cross-Sectional Design

Many important cross-sectional surveys are repeated at regular or irregular intervals so that estimates of changes can be made at the aggregate or population level. Examples include monthly labor force surveys, retail trade surveys, television and radio ratings surveys, and political opinion polls. These surveys are designed to give good estimates for the current population and the changes or movements that have occurred since the last survey or previous surveys. Typically surveys are conducted on a monthly, quarterly, or annual basis, although other intervals are possible, such as daily or weekly in the case of TV ratings and opinion polls. Surveys may also be conducted at longer intervals, such as 3 years, or repeated on an irregular basis, but in all cases there will be interest in estimating and analyzing changes at the population level and also various subgroups of the population, which are often defined geographically or in terms of sociodemographic variables.

Repeated cross-sectional surveys differ from longitudinal surveys, which are designed specifically to permit analysis of change at the individual or micro level and usually involve following an initial sample over several waves even if respondents move location. The need to follow respondents contributes to the cost and complexity of a longitudinal survey. In a longitudinal survey, there may be no interest in ensuring good cross-sectional estimates for each wave of the survey, and it may be difficult to do so. Longitudinal surveys are subject to attrition bias and conditioning effects but are valuable when the main aim of the survey is to understand changes at the individual level. In a repeated cross-sectional design, there is a strong emphasis placed on maintaining good sample representation to produce unbiased estimates for each time period. This can be done without following respondents over time.

In a repeated survey, an independent sample may be selected on each occasion, and so there will be essentially no overlap in the samples between time periods. There is then no possibility of conditioning effects or respondent fatigue, although there are the costs involved in making the initial contact with respondents and obtaining their cooperation. Valid estimates of changes at the population level can be calculated from independent samples. If y_t is the estimate for the population for time t and y_{t-s} the estimate for the population for time $t-s$, then the change or movement between the two time periods can be estimated by $y_t - y_{t-s}$. With independent samples, there will be differences between estimates for different periods because they are based on different samples. The sampling variance of the estimate of change will be the sum of the sampling variance on each of the estimates, so $Var(y_t - y_{t-s}) = Var(y_t) + Var(y_{t-s})$. If the sampling variances of each of the estimates are approximately equal, then the sampling variance of the estimate of change will be twice that of the cross-sectional estimates, and hence the standard error will be about 40% higher. Reliable estimates of changes can be obtained provided the sample sizes at each period are large enough and an efficient sample design is used, which produces unbiased estimates for each period. There is no need for the same sample size or design to be used at each occasion although it is usually efficient to do so.

An alternative design is to use the same sample on each occasion with some small changes to allow for new units in the population and remove units that are known to have left the population, where this knowledge is based on information that is sample independent. This design reduces costs, as it often is cheaper to survey people at the second and subsequent occasions, and also reduces the sampling variance of estimates of change because the effect of new units in the sample is minimized. The sampling variance is still present because respondents' characteristics may change, and the correlation between the values for the same respondent will be an important factor in determining the sampling variance.

Such a design will lead to respondents being included in the survey for a long time, which may lead to respondent fatigue and a reduction in both response rate and quality of the information reported. For these reasons, designs that involve some replacement or rotation of selected units are often used. Such designs are called *rotating panel designs* and can be set up in various ways so that there is a specified overlap in the sample between different periods and respondents are released from the survey after a specified time. Having overlap in the sample between consecutive surveys usually reduces the variance of the estimates of changes between consecutive periods, or any periods for which there is sample overlap. This is because $Var(y_t - y_{t-s}) = Var(y_t) + Var(y_{t-s}) - 2\sqrt{Var(y_t)Var(y_{t-s})}Corr(y_t, y_{t-s})$.

The correlation between the survey estimates, $Corr(y_t, y_{t-s})$, will be determined by the sample overlap, the sample design, and the correlation between individual-level values over the two time periods. It will have an appreciable beneficial effect if the sample overlap is high and the individual level correlation is also high and positive. If the individual-level correlation is low, then there is little benefit in having high sample overlap. Various rotation designs are possible and are chosen to balance cost, respondent burden, and impact on the important estimates of change.

In a rotating panel design, there is the opportunity to exploit the differences in the correlations between estimates that can be calculated from the overlapping and nonoverlapping samples to produce better estimates of level and change through composite estimation.

For a repeated survey, interest often will focus on the estimates of change between the most recent two periods, for example, between two consecutive months, $y_t - y_{t-1}$. However, to assess the general pattern of change, it is useful to estimate the changes over several periods to informally assess the trends in the time series of estimates. For a monthly survey, analysis may consider $y_t - y_{t-s}$ for $s = 1, 2, 3, 6, 12$. Formal methods of trend analysis can also be applied using filters or time series modeling. For monthly and quarterly surveys, seasonal adjustment also has to be considered.

In some applications the results from a repeated survey may be averaged. This may be done because the volatility of the estimates is too high, for example, for geographic areas. It may also be a deliberate part of the output strategy, for example, producing 3-month averages from a monthly survey. In this case, the positive correlation between estimates that is produced by having some sample overlap increases the sampling variances. For example, the average of 3 consecutive months would have variance

$$Var\left(\frac{y_{t+1} + y_t + y_{t-1}}{3}\right) = \frac{1}{9}\begin{bmatrix} Var(y_{t+1}) + Var(y_t) + Var(y_{t-1}) \\ + 2Cov(y_{t+1}, y_t) + 2Cov(y_t, y_{t-1}) \\ + 2Cov(y_{t+2}, y_{t-1}) \end{bmatrix}.$$

In a repeated cross-sectional design with independent samples, at each occasion within the 3-month period the covariances are all zero. If there is sample overlap, the covariance usually become positive and therefore increases the sampling variance. Simple moving averages are crude filters, and this feature affects trend estimates calculated by using filters. It would be better to use independent samples if averages over time are the main estimates of interest.

David Steel

See also Attrition; Composite Estimation; Cross-Sectional Survey Design; Longitudinal Studies; Panel Conditioning; Panel Surveys; Respondent Burden; Rolling Averages; Rotating Panel Design; Sampling Variance; Trend Analysis; Wave

Further Readings

Duncan, G. J., & Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, *55*, 97–117.

Kalton, G., & Citro, C. F. (1993). Panel surveys: Adding the fourth dimension. *Survey Methodology*, *19*, 205–215.

Kasprzyk, D., Duncan, G., Kalton, G., & Singh, M. P. (Eds.). (1989). *Panel surveys*. New York: Wiley.

McLaren, C. H., & Steel, D. G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. *Survey Methodology*, *26*, 163–172.

Steel, D. G. (2004). Sampling in time. In *Encyclopedia of social measurement* (pp. 823–828). San Diego, CA: Academic Press.

Steel, D. G., & McLaren, C. H. (2008). Design and inference over time. In C. R. Rao & D. Pfeffermann (Ed.), *Handbook of statistics: Vol. 29. Sampling theory methods and inference* (chap. 32). Amsterdam: North Holland.

Replacement

Replacement is a term used in two different contexts in surveys. With-replacement sampling refers to methods of sampling in which the unit selected in a particular draw is returned to the finite population and can be selected in other draws. The other context refers to the substitution (replacement) of a sampled unit with another unit as a result of difficulties in contacting or obtaining cooperation. The entry discusses both with-replacement sampling and replacement substitution.

With-Replacement Sampling

One approach to selecting an equal probability sample of n units from a finite population of N units is to draw one unit randomly from the N and then independently draw subsequent units until all n are selected. If the selected unit can be sampled more than once, then the method is *sampling with replacement*. This particular form of sampling is called *multinomial sampling*. There are many other ways of drawing a sample with replacement. For example, suppose the multinomial sampling procedure is used, but each unit is assigned its own probability of selection. If all the assigned probabilities of selection are not the same, then this is a simple way to draw a with-replacement, unequal probability sample.

When sampling with replacement, theoretically the same unit should be interviewed independently the number of times it is selected. Because this is operationally infeasible in most cases, the unit is interviewed once and the sampling weights are adjusted to account for the number of times it was sampled.

In practice, with-replacement sampling is not used frequently. However, a with-replacement sample does have some very important advantages, especially in the estimation of the precision of estimates in multistage samples. When a small fraction of the population units is sampled, it is often convenient to assume with-replacement sampling has been used when estimating variances. The variances of with-replacement and without-replacement samples are nearly equal when this condition exists, so the computationally simpler with-replacement variance estimator can be used.

Substitution

When data cannot be collected for a sampled unit, some surveys replace or substitute other units for the sampled ones to achieve the desired sample size (e.g., in Nielsen's television meter panels). Often, the substitutes are done during the field data collection; this is called *field substitution*, and the substituted unit may be called a reserve unit.

There are many different ways of selecting substitutes. Almost all substitution methods try to select the substitute from a set of units that match the characteristics of the nonresponding unit in some way. Some methods use probability mechanisms; for example, the units in the population that match the nonresponding unit are sampled with equal probability. Other methods are not based on probability mechanisms; for example, the interviewer is allowed to choose another household in the same block of the nonresponding household.

Replacement substitution tries to deal with unit nonresponse by replacing the nonresponding unit with another unit. In this sense, it is the equivalent of an imputation method. Just like imputation, substitution makes it difficult to accurately assess the statistical properties of estimates, such as their bias and variance. Substitute responses are typically treated as if they were the responses of the originally sampled units, which is usually not completely appropriate.

J. Michael Brick

See also Imputation; Nonresponse; Sampling Without Replacement

Further Readings

Kalton, G. (1983). *Introduction to survey sampling*. Newbury Park, CA: Sage.

Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole.

Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics*, *15*, 335–350.

Replicate Methods for Variance Estimation

Replicate methods for variance estimation are commonly used in large sample surveys with many variables. The procedure uses estimators computed on subsets of the sample, where subsets are selected in a way that reflects the sampling variability. Replication variance estimation is an appealing alternative to Taylor linearization variance estimation for nonlinear functions. Replicate methods have the advantage of transferring the complexity of variance estimation from data set end users to the statistician working on creating the output data set. By providing weights for each subset of the sample, called *replication weights*, end users can estimate the variance of a large variety of nonlinear estimators using standard weighted sums. Jackknife, balanced half-samples, and bootstrap methods are three main replication variance methods used in sample surveys. The basic procedure for constructing the replication variance estimator is the same for the three different methods. This entry describes the form of replication variance estimators, compares the three main approaches used in surveys, and ends with an illustration of the jackknife method.

Description

Replication variance methods involve selecting subsets from the original sample. Subsets can be created by removing units from the sample, as in the jackknife and balanced half-samples methods, or by resampling from the sample, as in the bootstrap method. A replicate of the estimator of interest is created for each subset. The replicate is typically constructed in the same way as the estimator of interest is constructed for the entire sample. Replication variance estimators are made by comparing the squared deviations of the replicates to the overall estimate. Thus, the replication variance estimator has the form $\sum_{k=1}^{L} c_k(\hat{\theta}^{(k)} - \hat{\theta})^2$, where k identifies the replicate, L is the number of replicates, $\hat{\theta}$ is the estimator of interest computed from the full survey, $\hat{\theta}^{(k)}$ is the k^{th} replicate of the estimator constructed using the k^{th} subset of the sample, and the c_k is the weight determined by the replication method and the survey design for the

k^{th} subset of the sample. The c_k's are often chosen so that the replication variance estimator is unbiased or nearly unbiased for the design variance of $\hat{\theta}$. The $\hat{\theta}$ in the replication variance estimator is often changed to the average of the replicates. Depending on the replication method, the number of replicates, L, may be large, but techniques for reducing L do exist.

Replication Method Comparison

For the jackknife method, subsets of the sample are constructed by removing elements systematically from the sample. The "delete-1" jackknife involves constructing n subsets where the i^{th} subset is the sample with element i removed. The jackknife approach works well, much like a Taylor linearization variance estimator, for sufficiently smooth functions. To relax the smoothness requirement, a "delete-d" jackknife can be used, where subsets with d elements deleted form replicates. The delete-d and even delete-1 approaches can generate a large number of replicates when the sample size is large. A "delete-a-group" jackknife, where subsets are created by dividing the sample into G groups and deleting one group at a time, reduces the number of replicates from n to G at the cost of degrees of freedom in the variance estimator.

The *half-samples method* was originally developed for the special case where the sample design is such that two elements within each stratum are sampled. Subsets of the sample are formed by removing one element from each stratum, so that the each subset is half of the original sample. If there are H strata, 2^H replicates could be created. Typically, the number of half-samples selected is reduced through a technique that maintains the variance of the stratified mean for the replication variance. The reduced set of samples has a property referred to as being balanced. The balanced half-samples approach, also called balanced repeated replication, can be used to estimate the variance for the Horvitz-Thompson quantile estimator unlike the delete-1 jackknife. The balanced half-samples method has been extended to stratified sampling, where more than two elements are selected per stratum.

The bootstrap method is widely used outside of survey statistics but has been adapted for unequal probability designs. For many bootstrap subset selection procedures, replicate samples are selected by sampling with replacement from the original sample. A large number of with-replacement samples are required, as there is additional variability due to the randomness of subset selection not present in the jackknife or balanced half-samples methods. The advantage of the bootstrap is that an estimate of the distribution of the estimator of interest is created by using the sampling distribution of the $\hat{\theta}^{(k)}$. Confidence intervals of size $1-\alpha$ can be constructed by taking the range of $\hat{\theta}^{(k)}$ after removing the largest and smallest $\alpha/2$ fraction of values. The bootstrap can be used for a wider variety of estimators than the delete-1 jackknife, such as standard quantile estimators. However, the bootstrap procedure generally requires more replicates than the alternative replication methods, and creating the bootstrap estimator for unequal probability samples is often more difficult than forming the jackknife.

For many survey estimators, the jackknife performs comparably to the Taylor linearization variance in terms of mean squared error. In simulation studies by Jun Shao and Donsheng Tu, as well as other authors, the jackknife tended to outperform balanced half-sampling when the jackknife was applicable. Simulation studies show for many estimators that the bootstrap is worse than either the jackknife or balanced half-samples. However, the bootstrap confidence intervals may outperform the jackknife confidence intervals in terms of coverage rates because the bootstrap intervals are not constrained to be symmetric.

Replicate Weight Example

Table 1 shows how replicate weights can be made in the case of two strata and a simple random sample within each stratum. The replicate weights are created by reallocating the original weight of the deleted element equally to the remaining elements in the same stratum. Suppose one quantity of interest is the ratio of the total of y, T_y, to the total of x, T_x. We will use the ratio of the Horvitz-Thompson estimators, $\hat{T}_y$ and $\hat{T}_x$, as the ratio estimator $\hat{R}$.

Associated with the replicate weights in the example is the vector of c_k. Each c_k depends on the stratum for which the k^{th} element belongs (see Table 2). Note that the average of the $\hat{T}_y^{(k)}$ (i.e., 4200/5) is equal to $\hat{T}_y$. The replication variance estimate of the Horvitz-Thompson estimator of T_y is 45150, which is the same as the Horvitz-Thompson stratified variance. The replicate variance for $\hat{R}$ is 0.0106. For comparison, the Taylor linearization variance for $\hat{R}$ is 0.0102.

The example illustrates how both the sample design and choice of using a delete-1 jackknife method impact the weights in the replication variance

Table 1 Example replication weights for a stratified sample

					Replicate				
Stratum	*Element*	*Original Weight*	*y*	*x*	*1*	*2*	*3*	*4*	*5*
1	1	21	9	10	0	42	21	21	21
	2	21	16	14	42	0	21	21	21
2	1	15	2	1	15	15	0	22.5	22.5
	2	15	14	13	15	15	22.5	0	22.5
	3	15	5	8	15	15	22.5	22.5	0

Table 2 Example jackknife weights and variance calculations

Replicate	c_k	$\hat{T}_y^{(k)}$	$\hat{T}_x^{(k)}$	$\bar{R}^{(k)}$	$c_k(\hat{T}_y^{(k)} - \hat{T}_y)^2$	$c_k(\hat{R}^{(k)} - \hat{R})^2$
1	20/42	987	918	1.08	10,290	0.0022
2	20/42	693	750	0.92	10,290	0.0033
3	28/45	952.5	976.5	0.98	7,875	0.0006
4	28/45	682.5	706.5	0.97	15,435	0.0011
5	28/45	885	819	1.08	1,260	0.0034
Sum		4,200	4,170		45,150	0.0106

estimator. In the case of multi-stage designs, care must be taken in selecting the subsets of the sample because of the correlation among elements in the same cluster. A common solution is to form replicate samples by selecting or deleting whole clusters; then estimated cluster totals are used in the variance estimator. Alternative replication procedures are available when the second-stage variance is important.

Jason C. Legg

See also Balanced Repeated Replication (BRR); Design-Based Estimation; Jackknife Variance Estimation; Taylor Series Linearization; Variance Estimation

Further Readings

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

McCarthy, P. J. (1969) Pseudo-replication: Half samples. *Review of the International Statistical Institute*, *37*, 239–264.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer.

REPLICATION

Replication is reanalysis of a study, building on a new data set that was constructed and statistically analyzed in the same way as the original work. Repeating the statistical analysis on the original data set is known as *verification* (or replication of the statistical analysis). Replicability should be maximized in both quantitative and qualitative works, as replication studies and verification of existing data sets may be extremely useful in evaluating the robustness of the original findings and in revealing new and interesting results.

Even if a given work will never actually be replicated, it still needs to be replicable, or else there is no possibility of refuting its findings; that is, it fails to hold the falsifiability criterion for scientific work. Because replicability is an underlying principle in science, most disciplines in social sciences hold some replication standards for publications, determining what information needs to be disclosed such that researchers may replicate the study without further guidance from the authors. In the case of survey research, the design and analysis of surveys call for many distinct decisions regarding the sampling, measurement, and methods used. As such, replication and verification of a survey is oftentimes extremely difficult to conduct, even by the original researchers. Thus, researchers should be sure to closely document their work process when gathering and analyzing their data.

To be able to replicate a given survey, researchers need to hold exact information on the sampling and the instruments used. First, decisions regarding the sample design and management need to be recorded, such as what list was used to sample from and how the sampling was conducted; who the interviewers (if any) were and how they were instructed and trained; what strata, quotas, or weights were used, if any; how many times people were recontacted; and how missing data was dealt with. Second, the exact instruments used should be recorded, including the question wording and order, split-half experimentation and any other randomizations (e.g., question order), counterbalancing or changes between questions, and whether the respondents were interviewed in their first language and what translations of the questionnaires were used, and so forth.

Another component of the work that needs to be replicable is the statistical analysis of the survey. In that vein, the researcher should document the construction of the variables, such as the coding of any open-ended questions (e.g., the list of categories, how many coders were employed, how they were instructed and trained, and what the intercoder reliability was), the construction of all variables used (e.g., the exact scales, techniques to deal with missing data, any rounding, mathematical transformations), as well as the exact software (statistical package and version) and statistical methods.

Even if replicable, studies in the social sciences usually cannot be entirely replicated when the measured phenomenon has changed between the original study and its replication attempt. This means that even if a researcher is able to retrieve the information from the real world and process and analyze it in the exact same way as did the original study, the results still may be different because the population under study had changed. Nevertheless, researchers should aim to maximize the replicability of their survey and analysis, and try to make certain that a full verification will be possible.

Pazit Ben-Nun

See also Coding; Missing Data; Sample Design; Sample Management

Further Readings

Herrnson, P. S. (1995). Replication, verification, secondary analysis, and data collection in political science. *PS: Political Science & Politics*, *28*(3), 452–455.

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Representative Sample

A representative sample is one that has strong external validity in relationship to the target population the sample is meant to represent. As such, the findings from the survey can be generalized with confidence to the population of interest. There are many factors that affect the representativeness of a sample, but traditionally attention has been paid mostly to issues related to sample design and coverage. More recently, concerns have extended to issues related to nonresponse.

Determining Representativeness

When using a sample survey to make inferences about the population from which the sampled elements were drawn, researchers must judge whether the sample is actually representative of the target population. The best way of ensuring a representative sample is to (a) have a complete list (i.e., sampling frame) of all elements in the population and know that each and every element (e.g., people or households) on the list has a nonzero chance (but not necessarily an equal chance) of being included in the sample; (b) use random selection to draw elements from the sampling frame into the sample; and (c) gather complete data from each and every sampled element. In most sample surveys, only the goal of random selection of elements is met. Complete and up-to-date lists of the populations of interest are rare. In addition, there are sometimes elements in the target population with a zero probability of selection. For example, in random-digit dialing telephone surveys, households without a telephone may belong to the population of interest, but if they do, then they have a zero chance of inclusion in the survey. Similarly, unless a cell phone frame is used in RDD sampling in addition to a landline frame, those with only cell phone service will have zero chance of inclusion. Thus, the random-digit dialing landline frame cannot fully represent the entire population of households. Researchers need to estimate sample coverage, which is an estimate of the proportion of elements in the population that are covered or included on the list or sample frame. To further complicate matters, almost all surveys have a significant number of sampled elements from which incomplete or no data are gathered because of unit nonresponse and item nonresponse.

Correcting for Biases

Given that two conditions of the criteria for a representative sample are rarely met in survey research, how is the likely representativeness of a sample determined? This is a crucial issue for sample surveys and one that is the subject of intense discussion and research. Representativeness is enhanced through one or more of the following. The first way is to rely on research conducted by other survey researchers on likely biases between the group of sampled elements (typically people) in a sample and the true characteristics of the population (i.e., population parameters, such as smoking prevalence or candidate preference). Much research has been conducted regarding the potential bias of working with an incomplete sampling frame of the population to draw the sample for the survey and into nonresponse of sampled elements (both item and unit survey nonresponse). Research regarding incomplete population coverage and nonresponse is often difficult to do because it is rare to have complete data on every element in a target population, as even censuses (such as the U.S. decennial census) have nonresponse and sample coverage problems. However, these data are the best available and are widely used as the "best guess" of the target population's characteristics. Most of the research on these two problems has found that nonresponse and sample frame noncoverage does bias the results of many sample surveys (thus lowering their external validity), as the responding sample often differs from characteristics of the entire population in nonnegligible ways. For example, for general population surveys in the United States and in many European countries, the responding sample is often better educated, more female, more likely to be home owners than renters, more white, and less ethnic than the general population of interest.

To correct for these biases, survey researchers invoke a second way to deal with these problems, that is, post-stratification. Post-stratification is the process of weighting some of the respondents or households in the responding sample relative to others so that the characteristics of the responding sample are essentially equal to those of the target population for those characteristics that can be controlled to census data (e.g., age, race, ethnicity, sex, education, and geography). By invoking post-stratification adjustments, the bias due to sample noncoverage and differential nonresponse theoretically is reduced.

The final correction in which researchers should engage is to limit the inferential population of a survey to those elements on the sampling frame with nonzero probability of inclusion. For example, a careful and conservative researcher who conducts a traditional landline random-digit dialing survey of adults in Georgia, would limit inferences to "adults living in households with landline telephones, who respond to surveys, in the state of Georgia." In practice this is rarely done because research sponsors typically want the survey to be representative of all adults in a given geopolitical area (e.g., Georgia), and too often empirical reports are written assuming (and implying) this is the case.

Future Research

Striving for representative samples is key when conducting sample survey research. However, it is important that consumers of survey-based information recognize that the standards for true representativeness are rarely met, but the biases produced by failures often are not severe enough to threaten the ultimate value of the survey findings. Because of these challenges, it is critical that research continue into the problems of sample design flaws of popular techniques (e.g., Internet surveys) and into the impact of unit and item nonresponse on findings from the survey.

Michael Edward Davern

See also Coverage Error; External Validity; Inference; Nonprobability Sample; Nonresponse Error; Population of Inference; Post-Stratification; Probability Sample; Sampling Error; Sampling Frame; Target Population

Further Readings

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Research Call Center

A research call center is the operational unit for survey data collection from which outbound telephone calls on a computer-assisted telephone interviewing survey are made. It can exist in many forms, including (a) a large single-site operation with many hundreds of booths; (b) several linked smaller sites; (c) a spare office with a few desks and phones; or (d) a virtual site, where the interviewers log into a Web application and do all the dialing from their homes.

Gathering data via the use of computer-assisted telephone interviewing is not the only research activity that might take place at the research call center, but this is the primary activity of such a set-up. Other research-related activities may include the following:

- Taking inbound calls, either as a support for an outbound survey (such as when a toll-free number is associated with a survey) or as part of data collection
- Editing and coding collected data
- Capturing (scanning or entering) data from hard-copy questionnaires
- Mailing out and receiving mail questionnaires
- Compiling the survey frame, including locating prior participants in a longitudinal study or recipients of a particular service being evaluated (i.e., by finding current addresses, telephone numbers, or both)
- Providing support services for other modes of data collection (e.g., providing a help desk function for respondents of a Web survey or collecting administrative data from field interviewers)

Research Versus Nonresearch Call Centers

Nonresearch call centers fall into two main groups: (1) predominantly outbound (such as a telemarketing or debt collection center) and (2) inbound (such as a customer assistance contact center for a bank or a catalogue sales support operation).

The common denominator among all call centers, including research call centers, is that there is a group of staff (interviewers or agents) sitting in booths either making or receiving calls. At this front stage of contact, there are rarely any formal educational requirements beyond high school reading ability; however, a clear voice is a necessity. Because calling volumes typically peak for only a small part of the day and often vary over the course of a year, most positions are part-time and often seasonal. As a result, many call centers in the same geographical locale tend to share the same labor pool, and the physical buildings in which they operate and the furniture needed tend to be very similar.

Technologically, there are a lot of similarities as well. All need a telephone system that can support many simultaneous calls and a computer system that will track and store the outcomes of the calls (such as completed questionnaires, completed applications for credit, or queries made and resolutions offered). However, an outbound center will require more sophisticated dialer equipment to place calls, whereas an inbound center will require a more sophisticated automatic call distributor and interactive voice response system to handle the queuing and directing of incoming calls to the most appropriate agent (not always an interviewer).

It is in the processes and procedures that the differences become more pronounced. For example, in comparing an outbound research survey operation and an outbound telemarketing operation, one of the main objectives of a research center is a high response rate, whereas for a telemarketing operation the overriding objective is to obtain a high volume of sales. Essentially, this is the difference between survey *quality* and telemarketing *quantity*, and this difference will play out in many ways. A research operation will make multiple calls to the same number following complex calling rules and will spend much more time on that one sample item. Shortcuts cannot be risked, the interviewers almost always will be paid by the hour rather than by the complete interview, in part on the assumption that this will help ensure they conduct the entire research task exactly as required (including gaining cooperation from as many respondents as possible and reading questions exactly as worded to minimize interviewer bias), and the dialer technology will be set to a slower rate to allow the interviewer time to read the call history notes from the previous call and to ensure that if the number answers, the interviewer is ready to take the call. The telemarketing operation will instead discard numbers very quickly and move onto fresh numbers, they will give their agents considerable latitude in the scripts and often pay commission rather than an hourly rate, and they will use high-volume predictive dialers to maximize the time their agents spend talking and selling as opposed to listening for answering machines or correctly classifying businesses.

Facilities

Research call centers are typically located in areas with a good supply of entry-level labor. Some are located near universities to take advantage of students looking for part-time work, while others are located in high immigrant areas if bilingual skills are needed. The long operating hours of call centers (usually 8:00 a.m. to 9:00 p.m. and often later if calls are being made across multiple Western time zones), combined with the usually low-income nature of the interviewing staff, often dictate the need for proximity to reliable and safe public transport.

Although a research call center can be constructed in any existing office, it usually consists of a large main floor where all the booths reside and has at least one training room and at least one break room, a secure insulated room to house the core technology, several offices to house administrative roles, and the usual amenities for any office building (bathrooms, storage, kitchenette, etc.).

The main floor is fitted out with both interviewer and supervisor booths. Because call centers usually operate for more than 80 hours a week, and the majority of staff will be working part-time, most booths are used by three or four interviewers each over the course of a working day or week. This limits the extent to which interviewers can personalize their booths.

The booths themselves usually are between 3 and 4 feet wide, with the trend being toward smaller ones matching the trend in the use of smaller hardware (flatscreen monitors, soft phones) and to maximize the use of the space available. Linear layouts (where there are straight rows of booths) offer the most efficient use of space and the best sight lines between interviewers and supervisors, but they might not be as attractive to work in as other configurations.

The more booths placed in an area, the more attention that needs to be paid to acoustics so that respondents cannot overhear other interviews taking place. Having good sound-absorbing surfaces (such as carpets, padded booth walls, and acoustic tiles on the ceiling) are essential, and if noise remains a problem, white noise generators can further assist.

Research Call Center Staffing

A typical research call center will have the following staff: interviewers, interview monitors, supervisors, and managers.

Interviewers

This is the entry level and most common position in a call center. Because the workload in most research call centers fluctuates considerably as projects come and go, and because of the predominance of household surveys where the most productive time to call is the relatively small window of weekday evenings, most if not all interviewing positions will be on a temporary and hourly or part-time basis. For a call center operating at full capacity, the number of interviewers on the payroll will need to be at least 2 to 3 times the number of interviewing booths available.

Few surveys require interviewers to have more than a high school education, although a clear speaking voice and the ability to persuade members of the public to take part in a survey also are essential. Most interviewers undergo approximately 40 hours of classroom training before they place their first call.

Interview Monitors

The key quality-control tool of a research call center is interview monitoring, where a trained member of the supervisory staff will listen in on interviews (either in real time while viewing simultaneously a copy of the interviewer's screen, or later to a recording of the interview along with a copy of the data collected) to check that the interviewer asked all questions in the prescribed way and faithfully recorded the answers given. Typically there will be one hour of monitoring for every 10 to 20 interviewer hours, so that 5% to 10% of all interviews are monitored.

Supervisors

Sometimes combined with the monitoring position, this position also acts as the first line of management and support for interviewers. Many supervisors are former interviewers. There will usually be at least one supervisor for every 10 to 20 interviewers. In larger operations, some supervisory positions are likely to be salaried, full-time positions. The supervisor level is usually the highest to which most interviewers can aspire without formal qualifications in survey methodology or a related field.

Managers

These positions usually carry both project management responsibility and day-to-day floor management. In larger operations these functions may be separated into specialist roles, but even where fewer than 50 booths are involved there will usually be two managers carrying both functions to provide redundancy. These are professional-level positions, requiring formal qualifications in survey methodology or a related field, along with substantial personnel management skills.

Large centers will also have specialized positions in technical support, human resource management, and training, whereas in smaller centers these functions will typically be spread among the supervisory and management staff.

Research Call Center Technology

The various technology components in a call center are, at the very minimum, the following:

- The telephone system, which includes the telephones on the interviewers' desks as well as the central PBX (private branch exchange) or call manager, which connects to the outside world
- The computer system, which runs the computer-assisted interviewing software and stores the collected data

Although research call centers often have a separate telephony infrastructure and computer infrastructure, as of 2007 these components are rapidly converging and there are already functions (such as dialing) that can reside in either. Many centers are also rapidly adopting the general call center trend toward Voice over Internet Protocol (VoIP) and other uses of Internet technology. VoIP allows the voice and data to be consolidated into a single network. This eliminates the need for parallel wiring systems (one for voice and another for data) within a center and reduces reliance on the more expensive telephone network.

Newer applications of computer-assisted interviewing are now written to be accessed via a Web browser, rather than residing on the interviewer's personal computer. This allows lower-specifications computers to be used in interviewing booths; saves the effort of individually configuring, testing, and managing all of the interviewer booths; and, combined with VoIP, allows interviewer stations to be set up rapidly and cheaply wherever a high-speed Internet connection is available, including in interviewers' homes.

Jenny Kelly

See also Calling Rules; Coding; Computer-Assisted Telephone Interviewing (CATI); Inbound Calling; Interactive Voice Response (IVR); Interviewer; Interviewer Monitoring; Outbound Calling; Predictive Dialing; Research Management; Supervisor; Voice over Internet Protocol (VoIP) and the Virtual Computer-Assisted Telephone Interview (CATI) Facility

Further Readings

Hansen, S. E. (2008). CATI sample management. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 340–358). New York: Wiley.

Kelly, J., Link, M., Petty, J., Hobson, K., & Cagney, P. (2008). Establishing a new survey research call center. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 317–339). New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Steve, K., Burks, A. T., Lavrakas, P. J., Brown, K., & Hoover, B. (2008). The development of a comprehensive behavioral-based system to monitor telephone interviewer performance. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 401–422). New York: Wiley.

Research Design

A research design is a general plan or strategy for conducting a research study to examine specific testable research questions of interest. The nature of the research questions and hypotheses, the variables involved, the sample of participants, the research settings, the data collection methods, and the data analysis methods are factors that contribute to the selection of the appropriate research design. Thus, a research design is the structure, or the blueprint, of research that guides the process of research from the formulation of the research questions and hypotheses to reporting the research findings. In designing any research study, the researcher should be familiar with the basic steps of the research process that guide all types of research designs. Also, the researcher should be familiar with a wide range of research designs in order to choose the most appropriate design to answer the research questions and hypotheses of interest.

Generally, the research designs can be classified into one of three broad categories based on the nature of research, purpose of research, research questions, sample selection, data collection methods, and data analysis techniques: (1) quantitative research designs, (2) qualitative research designs, and (3) mixed-research designs.

Quantitative Research Designs

Quantitative research is a deductive theory-based research process that focuses primarily on testing theories and specific research hypotheses that consider finding differences and relationships using numeric data and statistical methods to make specific conclusions about the phenomena. Quantitative research designs can be classified into one of four broad research design categories based on the strength of the research design's experimental control: (1) true experimental research designs, (2) quasi-experimental research designs, (3) pre-experimental research designs, and (4) nonexperimental research designs.

Although each of the categories of research design is important and can provide useful research findings, they differ in the nature of the evidence they provide in establishing causal relations between variables and drawing causal inferences from the research findings. Experimental designs are the most rigorous, powerful, and the strongest of the design categories to establish a cause–effect relationship. Nonexperimental designs are the weakest in terms of establishing a cause–effect relationship between variables because of the lack of control over the variables, conditions, and settings of the study.

True Experimental Research Designs

The true experiment is a type of research design where the researcher deliberately manipulates one or more independent variables (also called experimental variable or treatment conditions), randomly assigns individuals or objects to the experimental conditions (e.g., experimental or control groups) and controls other environmental and extraneous variables, and measures the effect of the independent variable on one or more dependent variables (experimental outcome). The experimental group is the group that receives the treatment, and the control group is the group that receives no treatment or sometimes a *placebo* (alternative treatment that has nothing to do with the experimental treatment). Thus, in a typical experimental study, the researcher randomly selects the participants and randomly assigns them to the experimental conditions (e.g., experimental and control), controls the extraneous variables that might have an effect on the outcome (dependent) variable, and measures the effect of the experimental treatment on the outcome at the conclusion of the experimental study.

It is important to emphasize that the experimental research design, if well conducted, is the most conclusive and powerful of all the research designs and the only research design that tests research questions and hypotheses to establish cause–effect relationships. For this reason it is sometimes called the "Golden Design."

The simple randomized experimental designs with two groups can be conducted using one of the following four basic experimental designs:

Randomized Two-Group Posttest-Only Designs

The two-group randomized experimental design involves two groups of individuals or objects which, ideally, are randomly selected from the population and which are randomly assigned to the experimental and control (comparison) groups (a single independent variable with two levels: experimental and control groups). The effects of the experimental treatment on the dependent variable (experimental outcome) are measured at the conclusion of the experiment. It is represented as

$$\text{Experimental Group} : R\ E\ O$$
$$\text{Control Group} : R\ C\ O$$

where R is random assignment of individuals or objects, E represents the experimental treatment, C represents the control condition (no treatment or placebo treatment), and O represents the posttest observation (measurement).

An example of this design would be testing an alternative wording of the mail survey cover letter, compared to a standard cover letter, to determine whether the new cover letter raised cooperation. Households would be randomly assigned to either the standard or alternative cover letter. The resulting response rates between the two groups would represent the dependent variable used to test the hypothesis that the new wording raises cooperation.

Randomized Two-Group Pretest–Posttest Designs

This experimental design involves two groups of individuals or objects randomly selected from the population and randomly assigned to the experimental and control groups (a single independent variable with two levels: experimental and control groups). The two groups are pretested on the dependent variable before administering the experimental treatment and posttested on the same dependent variable at the conclusion of the experiment. This design is represented as

$$\text{Experimental Group} : R\ O_1\ E\ O_2$$
$$\text{Control Group} : R\ O_1\ C\ O_2$$

where R is random assignment of individuals or objects, E represents the experimental treatment, and C represents the control condition (no treatment or placebo treatment). The O_1 represents the pretest observation (measurement), and the O_2 represents the posttest observation (measurement).

An example of this design would be a telephone survey questionnaire that measures the effects of new information on approval versus disapproval of a proposed city bond to fund the building of a new bridge. All respondents would be asked whether they favor or oppose the new bridge funding early in the questionnaire. Later in the questionnaire they would be asked the same favor–oppose question again, but a random half of them would first be told some information about the value of the new bridge and the other half would not be told this information. Nothing else in the questionnaire would change. The difference in answers between the before- and after-questions about the funding for the two groups would serve as the dependent variable to test the hypothesis that the new information raises support for the bridge funding.

Solomon Four-Group Designs

This experimental design is a combination of the randomized two-group posttest-only design and the randomized two-group pretest–posttest designs. It involves randomly selecting a sample of subjects from the targeted population and randomly assigning the random sample to one of four groups. Two of the groups are pretested (Experimental and Control Groups 1) and the other two are not (Experimental and Control Groups 2). One of the pretested groups and one of the not pretested groups receive the experimental treatment. All four groups are posttested on the dependent variable (experimental outcome). The design is represented as

$$\text{Experimental Group 1} : R\ O_1\ E\ O_2$$
$$\text{Control Group 1} : R\ O_1\ C\ O_2$$
$$\text{Experimental Group 2} : R\ E\ O_2$$
$$\text{Control Group 2} : R\ C\ O_2$$

Here, the researcher has two independent variables with two levels. One independent variable is the experimental conditions with two levels (experimental and control groups), and the other independent variable is the pretesting condition with two levels (pretested and not pretested groups). The value of this design is that it allows the researcher to determine if the pretest (O_1) has an effect on the resulting answer given in the posttest.

An example of this design would be one that builds on the previous example of the experiment to test the effect of the information about the value of the new bridge. However, in the Solomon four-group design, there would be two more randomly assigned groups of respondents, ones who were not asked whether they favored or opposed the bridge funding at the beginning of the questionnaire. Instead, one of these groups would be the second control group, asked only their opinions about the bridge funding later in the questionnaire. The other group would be the second experimental group, asked their opinions about the bridge funding only later in the questionnaire but after first being given the information about the value of the bridge. This design would allow the researchers to test not only the effects of the information but also whether the saliency of the bridge funding, by asking about it first before giving the new information, affected opinions given later about the funding.

Experimental Factorial Designs

Experimental factorial designs are extensions of single independent variable experimental designs to situations where there are two or more independent variables that are controlled by the researcher. Factorial designs allow the researcher to examine simultaneously the effects of one or more independent variables individually on the dependent variable (experimental outcome) as well as their interactions. These interactions cannot be examined by using single independent variable experimental designs.

The term *factorial* refers to experimental designs with more than one independent variable (factor). Many different experimental factorial designs can be formulated depending on the number of the independent variables. The Solomon four-group design is an example of a 2×2 factorial design with treatment conditions (treatment and control groups) crossed with pretesting conditions (pretested and not pretested groups).

Quasi-Experimental Research Designs

Quasi-experimental research is used in situations where it is not feasible or practical to use a true experimental design because the individual subjects are already in intact groups (e.g., organizations, departments, classrooms, schools, institutions). In these situations it is often impossible to randomly assign individual subjects to experimental and control groups. Thus, quasi-experimental designs are similar to experimental designs in terms of one or more independent (experimental) variables being manipulated, except for the lack of random assignment of individual subjects to the experimental conditions (i.e., experimental and control groups). Instead, the intact groups are assigned in a nonrandom fashion to the conditions. Types of quasi-experimental designs include nonequivalent control group designs, longitudinal research designs, and multi-level research designs.

Nonequivalent Control Group Design

The nonequivalent control group design involves assignment of intact nonequivalent groups (e.g., classrooms, schools, departments, and organizations) to experimental conditions (experimental and control groups). Thus, the intact groups are assigned to the treatment conditions and not the individual subjects, as was the case in the true experimental designs. For example, in a study of the effects of a new curriculum of students' knowledge of science and attitudes toward science, some classrooms would be assigned to receive the new curriculum and others would not. Toward the end of the school year, all students are measured on their science knowledge and attitudes toward science. Because the effects are being measured at the level of the individual student, but the students themselves were not randomly assigned to the control and treatment condition, this is a quasi-experiment, not a true experiment.

Longitudinal Research Designs

Longitudinal, repeated-measures, or time-series research designs involve repeated measurement or observation on the same individuals at several points over a period of time. It is an elaboration of the one-group pretest–posttest design and focuses primarily on change, growth, and developmental types of research questions across many different disciplines such as medicine, public health, business, and social and behavioral sciences. Longitudinal designs, if well designed and conducted, are usually more complex, time consuming, and expensive than the other types of research designs.

Multi-Level Research Designs

Multi-level or hierarchical research designs involve the nesting of individuals (micro-level units) within organizations (macro-level units) and having explanatory independent variables characterizing and describing both levels. For example, in a two-level design, the emphasis is on how to model the effects of explanatory variables (predictors) at one level on the relationships occurring at another level. These multi-level and hierarchical structured data present analytical challenges that cannot be handled by traditional linear regression methods because there is a regression model for each level of the hierarchy. Thus, hierarchical models explicitly model the micro and macro levels in the hierarchy by taking into consideration the interdependence of individuals within the groups.

Pre-Experimental Research Designs

Pre-experimental research designs are simple designs with no control groups. These designs are questionable because they lack control and thus should be used for exploratory or preliminary examination of research problems.

One-Group Posttest Experimental Design

The one-group experimental design, also called the one-shot experimental design, takes a single group of subjects or objects exposed to a treatment (X) and observes and measures its effects on the outcome (O). This simple design is represented as

$$X \rightarrow O$$

This is the most basic and simple design in experimental research. It is used as a starting point for preliminary examination of the precausal relationship of research problems for the purpose of developing better-controlled future experimental designs.

One-Group Pretest–Posttest Design

The one-group pretest–posttest design involves a single group of individuals or objects that are pretested or measured (O_1), exposed to an experimental treatment (X), and posttested or measured (O_2). This design is represented as

$$O_1 \rightarrow X \rightarrow O_2$$

Nonexperimental Research Designs

Nonexperimental or descriptive research designs aim to answer research questions about the current state of affairs, identify factors and relationships among them, and create a detailed quantitative description of phenomena. Thus, it provides a snapshot of the feelings, opinions, practices, thoughts, preferences, attitudes, or behaviors of a sample of people, as they exist at a given time and a given place. For example, measuring the attitudes of the employees in the organization toward adapting new technologies is an example of a research question that can be carried on using a nonexperimental descriptive survey research design. The following are short descriptions of some of these designs.

Nonexperimental Survey Research

Survey research is a systematic research method for collecting data from a representative sample of individuals using instruments composed of closed-ended and/or open-ended questions, observations, and interviews. It is one of the most widely used nonexperimental research designs across disciplines to collect large amounts of survey data from a representative sample of individuals sampled from the targeted population using a variety of modes such as face-to-face, telephone, mail, and electronic (Web-based and email). Each of these data collection modes has its own advantages and disadvantages in terms of cost, duration, and response rate. Thus, the key goal of nonexperimental survey research is to collect data and describe the behaviors, thoughts, and attitudes of a representative sample of individuals at a given point in time and place.

Survey research is considered one of the most important research designs, and survey instruments and survey methods are frequently used to collect data for the other quantitative, qualitative, and mixed research designs. For example, it can be used to collect data for correlational research studies, experimental studies, and quasi-experimental studies.

Correlational Research

Correlational research is a type of descriptive nonexperimental research because it describes and assesses the magnitude and degree of an existing relationship between two or more continuous quantitative variables with interval or ratio types of measurements or discrete variables with ordinal or nominal type of measurements. Thus, correlational research involves collecting data from a sample of individuals or objects to determine the degree of the relationships between two or more variables for the possibility to make predictions based on these relationships. There are many different methods for calculating a correlation coefficient, which depends on the metric of data for each of the variables. The most common statistic that measures the degree of the relationship between a pair of continuous quantitative variables, having interval and ratio types of measurements, is the Pearson product–moment correlation coefficient, which is represented by the letter r.

Alternative correlation coefficients can be used when the pair of variables has nominal or ordinal types of measurement. If the pair of variables is dichotomous (a nominal type of measurement having only two categories), the Phi coefficient should be used. If the pair of variables has ordinal type of measurement, the Spearman rank order correlation coefficient should be used.

Another type of correlational research involves predicting one or more continuous quantitative dependent variables from one or more continuous quantitative independent variables. The most common statistical methods used for prediction purposes are simple and multiple regression analyses.

The significance of correlational research stems from the fact that many complex and sophisticated statistical analyses are based on correlational data. For example, logistic regression analysis and discriminant function analysis are quite similar to simple and multiple regression analyses with the exception that the dependent (criterion) variable is categorical and not continuous as in simple and multiple regression analyses. Canonical analysis is another statistical method that examines the relationship between a set of predictor (independent) variables and a set of criterion (dependent) variables. Path analysis and structural equation modeling are other complex statistical methods that are based on correlational data to examine the relationships among more than two variables and constructs.

Causal-Comparative Research

Causal-comparative or ex post facto research is a type of descriptive nonexperimental research because it describes the state of existing differences among groups of individuals or objects as they existed at a given time and place and attempts to determine the possible causes or reasons for the existing differences. Thus, the basic causal-comparative approach starts with selecting two or more groups with existing differences and comparing them on an outcome (dependent) variable. Also, it attempts to examine and explain the possible causes of the existing differences between the groups.

Some causal-comparative designs involve only two independent groups to be compared on a particular continuous dependent variable, for example, studying the differences between boys and girls on math achievement. In this causal-comparative study, the researcher needs to analyze the collected data using *t*-test for testing the research hypothesis that there are differences between the two independent sample means. Some other causal-comparative research designs involve more than two groups, for example, studying differences between white, black, and Hispanic students on math achievement. In this study, the researcher needs to use analysis of variance (ANOVA) to analyze the data.

Other causal-comparative designs involve studying differences between (among) two or more independent groups on two or more related dependent variables. In this case, multivariate analysis of variance (MANOVA) statistical procedure should be used to analyze the data to determine whether two or more independent groups differ on more than a single dependent variable.

It is important to note that the *t*-test, ANOVA, and MANOVA are parametric statistical procedures that require interval- or ratio-level data, a large sample size, and meeting the requirements of statistical assumptions (e.g., normality, independence of observations). The nonparametric counterparts for these statistical methods should be used with nominal- or ordinal-level data and when one or more of the assumptions are violated in the research study and when the sample size is small. For example, a nonparametric statistical method such as Mann–Whitney *U* is an alternative to the parametric *t*-test.

Meta-Analysis Research

The meta-analysis design is used to quantitatively and systematically summarize and synthesize the research results and findings from a collection of primary studies that address and test the same research question. Meta-analytic research methods have established five major general stages that guide meta-analysts in their systematic quantitative review. These stages include (1) formulating research problems, (2) collecting primary research studies, (3) evaluating primary studies, (4) analyzing and modeling the meta-analytic data, and (5) interpreting and presenting the meta-analytic results.

Generally, the key goals of meta-analysis methods are to (a) produce quantitative summary measures of the effect sizes, (b) assess the heterogeneity (variation) among the effect sizes, and (c) model and explain the heterogeneity between the effect sizes using known study and sample characteristics as exploratory variables in the specified meta-analytic regression model.

Qualitative Research Designs

Qualitative research is inductive and context-specific research that focuses on observing and describing a specific phenomenon, behavior, opinions, and events that exist to generate new research hypotheses and theories. The goals of qualitative research are to provide a detailed narrative description and holistic interpretation that captures the richness and complexity of behaviors, experiences, and events in natural settings. Thus, qualitative research is an inductive research process, logically emerging from the specific phenomena to general conclusions and theories about the phenomena based on data collected by observations, documents, physical artifacts, interviews, and focus groups.

Case Study

Case study is an in-depth examination and intensive description of a single individual, group, and organization based on collected information from a variety of sources, such as observations, interviews, documents, participant observation, and archival records. The goal of the case study is to provide a detailed and comprehensive description, in narrative form, of the case being studied.

Ethnographic Research

Ethnographic research is a qualitative research design that is used for studying social groups, cultures, and human interactions in natural cultural and social settings. The goal of the ethnographic study is to provide a detailed, in-depth, and holistic narrative description of the group and the cultural setting being studied. The primary ethnographic data collection methods are in-depth interviews and participant observation to comprehensively describe a cultural and social setting.

Phenomenological Research

Phenomenological research, or phenomenology, is a qualitative research method in which the researcher attempts to understand and explain how an individual or a group of individuals experience a particular phenomenon from the individual's or individuals' own perspective(s). The primary method of data collection used in phenomenology is in-depth interviews of individuals who have experienced the phenomenon.

Action Research

Action research is a systematic research inquiry conducted by teachers, principals, school counselors, managers, or any other educational or organizational practitioners in the educational and organizational setting to collect information about educational and organizational practices and operations to resolve matters of concern or a problem in a particular setting such as classroom, playground, library, department, or company. Simply stated, action research is a study conducted by educational and organizational practitioners to help them to develop alternative reflective practices that lead to positive changes within their educational and organizational settings.

Historical Research

Historical research is a systematic process for searching, exploring, summarizing, and reporting past information and events using primary and secondary sources of historical data to gain understanding of historical events, issues, and policies. Primary sources of historical data are the original firsthand artifacts, documents, observations, oral presentations, diaries, photographs, and audio-visual recordings of past events. Secondary sources are secondhand nondirect oral and written documentations of past events that are summarized and documented by others and not the original primary sources.

Grounded Theory Research

Grounded theory research is an inductive qualitative research design that is used for generating and developing theories and explanations based on systematically collected qualitative data. The data collection process in grounded theory research is usually an ongoing iterative process that starts with collecting and analyzing qualitative data that leads to tentative theory development. Then, more qualitative data are collected and analyzed that lead to further clarification and development of the theory. The qualitative data collection and further theory development process continues until the particular theory is developed that is "grounded" in the data.

Mixed-Methods Research Designs

Mixed-methods research designs involve research studies that employ both quantitative and qualitative research methodologies to address the proposed research questions. Thus, mixed research methods combine the deductive and inductive inquiries of the scientific research methods as well as use a variety of data collection and analysis methods. The quantitative and qualitative methods can be conducted concurrently or sequentially to address a research question or questions. The mixed-methods research designs require from the researcher a considerable amount of time and energy as well as training in both quantitative and qualitative research designs. However, one of the significant advantages of the mixed-methods research design is that it provides a more comprehensive and enhanced image of the research problem that is under investigation than would either one of the designs (quantitative or qualitative) by itself. Specifically, the mixed-methods research designs can be classified into three types: exploratory, explanatory, and triangulation.

Exploratory Mixed-Methods Research Designs

Using this design, the researcher first conceptualizes a qualitative research study. Second, the researcher collects and analyzes the qualitative data. Third, the researcher uses the findings from the qualitative data analysis to conceptualize a quantitative research study. Finally, the researcher collects and analyzes the quantitative data to validate the qualitative findings.

Explanatory Mixed-Methods Research Designs

Using this design, the researcher first conceptualizes a quantitative research study. Second, the researcher collects and analyzes the quantitative data. Third, the researcher conceptualizes a qualitative research study. Finally, the researcher collects and analyzes the collected qualitative data to clarify and enhance the quantitative research findings.

Triangulation Mixed-Methods Designs

Using this design, the researcher simultaneously conceptualizes quantitative and qualitative research studies. Then, the researcher simultaneously collects and analyzes both quantitative and qualitative data. Finally, the researcher uses the results from the quantitative and qualitative studies to validate findings from both studies.

Sema A. Kalaian

See also Dependent Variable; Experimental Design; Factorial Design; Independent Variable; Longitudinal Studies; Random Assignment; Trend Analysis

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
Frankel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill.
Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. Hoboken, NJ: Wiley.
Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Research Hypothesis

A research hypothesis is a specific, clear, and testable proposition or predictive statement about the possible outcome of a scientific research study based on a particular property of a population, such as presumed differences between groups on a particular variable or relationships between variables. Specifying the research hypotheses is one of the most important steps in planning a scientific quantitative research study. A quantitative researcher usually states an a priori expectation about the results of the study in one or more research hypotheses before conducting the study, because the design of the research study and the planned research design often is determined by the stated hypotheses. Thus, one of the advantages of stating a research hypothesis is that it requires the researcher to fully think through what the research question implies, what measurements and variables are involved, and what statistical methods should be used to analyze the data. In other words, every step of the research process is guided by the stated research questions and hypotheses, including the sample of participants, research design, data collection methods, measuring instruments, data analysis methods, possible results, and possible conclusions.

The research hypotheses are usually derived from the stated research questions and the problems being investigated. After the research hypotheses are stated, inferential statistical methods are used to test these hypotheses to answer the research questions and make conclusions regarding the research problems. Generally, in quantitative research designs, hypothesis testing and the use of inferential statistical methods begin with the development of specific research hypotheses that are derived from the study research questions. Research hypotheses differ from research questions in that hypotheses are specific statements in terms of the anticipated differences and relationships, which are based on theory or other logical reasoning and which can be tested using statistical tests developed for testing the specific hypotheses.

The following are two examples of research questions and possible corresponding research hypotheses.

The first example is a nondirectional hypothesis, whereas the second is a directional hypothesis:

> *Research Question 1:* What differences exist in attitudes toward statistics between male and female Ph.D. candidates in the technology program?
>
> *Research Hypothesis 1:* There are statistically significant differences in attitudes toward statistics between male and female Ph.D. candidates in the technology program.
>
> *Research Question 2:* Is rapport with graduate students different for professors using a student-centered teaching method than for those using a teacher-centered teaching method?
>
> *Research Hypothesis 2:* Professors who use a student-centered teaching method will have a greater positive rapport with their graduate students than professors who use a teacher-centered teaching method.

It is important to note that in using some types of research designs, it is unnecessary and inappropriate to state research hypotheses because in these types of studies, it would be impossible to anticipate or predict the possible outcomes and findings of the study. For example, survey research that is designed to explore and describe the characteristics (e.g., attitudes, opinions) of a particular population often proceeds with no need to state research hypotheses. In addition, in qualitative research designs such as ethnographic, case studies, grounded theory, and phenomenological research, research hypotheses usually are not formulated at the beginning of the research. They are usually generated and emerged as qualitative data are collected and more understanding is gained about the phenomenon under investigation and may lead to follow-up quantitative studies to investigate the newly formed hypotheses.

There are three distinct properties that are shared by all types of research hypotheses. The first property is whether the hypothesis represents a difference between groups or a relationship hypothesis. The second property of research hypotheses is whether it is the null hypothesis or an alternative hypothesis. The third property is whether the hypothesis is directional or nondirectional.

Difference-Between-Groups Hypotheses Versus Relationship Hypotheses

Research hypotheses can take many different forms depending on the research design. Some hypotheses may describe and examine the relationship between two variables. Some hypotheses may examine the differences between two or more groups. Other hypotheses may examine the effect of particular explanatory independent variables on the dependent outcome variable.

Nevertheless, research hypotheses can be classified into two broad categories:

> *Difference hypotheses.* This type of research hypothesis is used for group comparison purposes in randomized experimental designs, quasi-experimental designs, and causal-comparative research designs. Research hypotheses 1 and 2, presented in the previous section, are difference hypotheses.
>
> *Relationship hypotheses.* This type of research hypothesis is used to examine the relationships between two or more variables in correlational research designs. The following is an example of a relationship research hypothesis: There is a positive (direct) relationship between Ph.D. students' attitudes toward statistics and their achievement in the research methods courses.

Null Versus Alternative Hypotheses

In hypothesis testing, there are two kinds of research hypotheses. One is the null hypothesis, symbolically stated as H_0, and the other is the alternative hypothesis, symbolically stated as H_1. The null hypothesis always states that there are no differences between groups on a particular variable being studied, no effects of particular independent variables on the dependent outcome variable, or no relationship between the variables being examined. In contrast, the alternative hypothesis states there are differences between two or more groups, there is an effect of an independent variable on the dependent variable, or there are relationships between pairs of variables.

It is important to note that the alternative hypothesis represents the researcher expectations about the range of the possible values that the hypothesized parameter might take in the population. For this reason it is impossible to test the alternative hypothesis directly for all the possible parameter values in that range. To confirm or disconfirm the research hypothesis, the researcher usually devises and tests the null hypothesis. The null hypothesis can be thought of as the complement of the alternative hypothesis where the hypothesized parameter is equated to a single value that can be directly tested by a statistical test. Rejecting the null hypothesis is an indirect way of "confirming" (supporting) the researcher's alternative hypothesis.

It is equally important to note that based on the results of the test statistic, the null hypothesis is either rejected or not rejected, but it can never be accepted. For example, if the premise that there is no difference between groups (i.e., the null hypothesis) is rejected, then the researcher concludes that there is a significant difference between the groups. On the other hand, if the same null hypothesis is not rejected, then the researcher concludes that there is no significant detectable difference between the groups of the study (at least not as measured in the current study).

Directional Versus Nondirectional Hypotheses

In hypothesis testing, it is important to distinguish between directional hypotheses and nondirectional hypotheses. The researcher should make the choice between stating and testing directional or nondirectional hypotheses depending on the amount of information or knowledge the researcher has about the groups and the variables under investigation. Generally, a directional hypothesis is based on more informed reasoning (e.g., past research) than when only a nondirectional hypothesis is ventured. Research Hypothesis 1 stated earlier is an example of a nondirectional research hypothesis. It can be represented along with its null hypothesis symbolically as

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2.$$

μ_1 and μ_2 are the mean attitudes toward statistics held by the male and female Ph.D. students, respectively.

Research hypothesis 2 stated earlier is an example of a directional research hypothesis and it can be represented along with its null hypothesis symbolically as

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2.$$

μ_1 and μ_2 are the mean levels of rapport with their students for professors who use a student-centered method and those who use a teacher-centered teaching method, respectively.

Sema A. Kalaian and Rafa M. Kasim

See also Alpha, Significance Level of Test; Alternative Hypothesis; Dependent Variable; Independent Variable; Null Hypothesis; Research Design; Research Question

Further Readings

Frankel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill.

Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology.* New York: Wiley.

Research Management

The function of research management is to coordinate disparate activities involved in planning and executing any research project. Ability to grasp the big picture, coupled with attention to detail, is required in this function. Regardless of research team size or the size of the project, the same management activities are required.

Research Plan

Research management begins at the initial phase of the research process. As survey research texts point out, identification of the research problem is the first step. From this, the research team develops testable hypotheses and the research plan. This research plan or process is the cornerstone of research management, and the more detailed the plan is, the more efficient the project implementation phase will be.

Identification of Sample

Once the research team has identified the population of interest, it is time to address sampling issues. Sample composition required to meet project goals, sample size (based on power calculations and tempered by monetary considerations), and source of the sample are specified in the research plan.

Timeline

Effective research management requires the development of a realistic timeline to ensure that the project proceeds in an organized and logical manner. Recognizing that the research process seldom proceeds on schedule, additional time is included in the timeline so that a delay does not compromise the entire project.

Funding

Research management requires an understanding of accounting, budgeting, and some basic contract management. An important research management function is to secure and manage funds necessary for the project's success. After the research plan is developed, the research budget is finalized. The budget must include all costs required to execute the project in accordance with the research plan. Survey administration modes impact the budget differently. Traditionally, mail has been the least expensive, followed by Internet, Web, computer-assisted self-interviewing (audio or not), and computer-assisted telephone interviewing (CATI). Face-to-face interviewing, including computer-assisted personal interviewing (CAPI), is the most expensive mode.

Often, a research proposal is submitted to federal, state, university, corporate, or other funding agencies to obtain adequate project funds. When working with a team of researchers, this activity can, and probably should, be shared. A first step in the proposal process is to identify an appropriate funding entity, that is, an agency where agency goals and those of the research project are similar. It also is important that the funding agency has the capacity to adequately fund the project. Conducting an expensive research project may require securing money from several funding agencies. Be aware that when obtaining funding, the funding agency often specifies question topics for the survey instrument and this may increase the questionnaire length and thus the cost.

Obtaining adequate funds impacts the timeline because of agency funding cycles and proposal requirements. It is not unusual to submit a proposal and revisions multiple times before funding is secured; thus, additional time should be included in the timeline for such a contingency. At the time funding is awarded, the research team and funding agency should have a shared understanding, in writing, about project deliverables.

Proposals submitted to organizations that are not federal, state, or university entities and do not receive funding from these sources are developed to address the client's requirements. Typically these proposals and the timeline for developing them are very short; therefore, it is prudent to have boilerplate proposal language developed to reduce proposal preparation response time. In most cases, the client will have specified the project budget. Research firms submitting successful proposals must either meet all proposal requirements within the specified budget or convince the client that the proposed activities cannot be performed with the budget specified.

Questionnaire Development

Research management's role in questionnaire development is to ensure that included survey questions answer the research questions and are appropriate for the selected survey mode. Competent research management occasionally requires interjecting realism into the questionnaire development process. Occasionally, those involved in writing survey questions assume that potential respondents have similar subject matter knowledge and language comprehension levels as themselves. When necessary, it is the responsibility of research management to disabuse them of this belief.

Unique challenges are encountered when sampling a multi-cultural, multi-lingual population because without interviewing in multiple languages, researchers cannot extrapolate their findings to the entire population. At the beginning of such projects, it is important to secure the services of an experienced translator or translation company. Researchers may be tempted to economize on the translation and discover, after the fact, that simply speaking and reading a language does not mean that the person is capable of translating a questionnaire. Specialized subject-matter knowledge greatly aids in the translation process. Cultural competencies and sensitivities also affect the translation. After the initial translation is completed, the questionnaire often is back-translated to ensure accuracy of the original translation.

Protection of Human Subjects

Once the research plan is completed, the sample defined, the questionnaire developed and translated, and federal, state, or university funding secured, the project must be submitted to a Committee for the Protection of Human Subjects. When only one funding agency is involved, the project is submitted to that agency's approved committee. In executing their mandate to protect research participants from harm, the committee may require any number of changes to a project. Mounting a large and expensive project involving multiple funding agencies requires submitting the project to multiple human subject committees. Adequate time should be included in the project timeline to account for differences in the frequency with which the committees meet and the necessity of coordinating changes required by each committee into a cohesive project approved by all.

Projects that do not involve federal, state, or university funds are subject to the review requirements of the client. Some commercial organizations have an institutional review board that performs the functions of the Committee for the Protection of Human Subjects, and projects must be approved through this committee. When working with clients whose organizations do not require the approval of an institutional review board, it is prudent to obtain approval, in writing, for all aspects of the project prior to data collection.

Data Collection

Activities involved in data collection are dependent upon the mode of administration and whether the research team is actively involved in data collection or is coordinating this activity with another entity. When collecting data in-house, research management's role is to ensure that the data collection team is large enough to complete the project on schedule and within budget and that the interviewers are adequately trained and supervised. If data collection is outsourced, research management involves identifying an organization capable of performing the required activities at the level of quality that is required. Active oversight of the contractor is another function of research management. Oversight involves ensuring that data are being collected according to the research plan; that, if required, the contractor adheres to the regulations of the Committee for Protection of Human Subjects; and that respondent confidentiality is not compromised.

When collected via the Internet, Web, CAPI, CATI, and other computer-assisted methods, data automatically are captured electronically, thus eliminating additional data entry costs. Projects conducted via mail or other paper-and-pencil modes require data entry; therefore, provisions should be made to test a proportion of the records for accuracy of data entry. For qualitative research, or when clients are uncertain about the appropriate response categories, open-ended text questions may be included on the survey instrument. While clients often see this as a way of obtaining the "best" data from respondents, question administration and coding present unique challenges. Frequently respondents have not given much thought to issues addressed in open-ended questions and, therefore, may provide incomplete, somewhat incoherent, or other unusual responses. Prior to coding open-ended responses, it is suggested that clients be provided with the text responses and help develop a coding scheme for the research organization. In doing so, the client will gain the greatest benefit from the open-ended question responses. When clients are reluctant to do this, those coding the text responses should look for common response themes, receive client approval for the coding scheme, and code the data accordingly.

Data obtained using computer-assisted modes are relatively clean when the software programming is thoroughly checked for accuracy, response categories are appropriate, and when incomplete or out-of-range responses are not allowed. Paper-and-pencil administration often results in incomplete data because of item nonresponse, out-of-range responses, and illegible responses. Data cleaning becomes more onerous in this situation, and research management's role is to ensure that data quality is not compromised. Decisions regarding data cleaning, recoding, and acceptable data quality should be delineated in the data set documentation.

Data Analysis

Before data analysis and report writing, the audience for the report needs to be determined. The client should be asked about the degree of complexity desired and the type of report needed. While it may be tempting to "dazzle" a client by providing high-level statistical analysis in reports, if the client is unable to use the analysis, then the report may not be useful. All specifications of this deliverable should be delineated in writing prior to the start of analysis. A function of project management is to ensure that clients receive usable deliverables that meet their needs and, at the same time, enhance the survey organization's reputation.

Data Set and Documentation Development

The type of data set developed (e.g., SAS, SPSS) is dependent upon project deliverables. Documentation makes the data set useful to data users. In addition to a codebook, documentation usually includes a description of the sampling strategy used; source of the sample; information about the data collection process;

measures of data quality; and data weighting, if appropriate. Project complexity affects data set documentation development and can take weeks or months to complete. Just as with data collection, responsibility for overseeing this activity is a function of research management.

Project Success

A thorough understanding of the research process, coupled with a basic knowledge of fiscal management, is required for competent project management. Ability to communicate both orally and in written form with researchers, support staff, and interviewers is vital to project success. While not required, being personable and having some tact makes communication with the research team and others easier. Although a research team usually consists of many persons, ultimately research project success is dependent upon competent research management.

Bonnie D. Davis

See also Audio Computer-Assisted Self-Interviewing (ACASI); Codebook; Coding; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Interviewing; Protection of Human Subjects; Questionnaire; Research Design; Survey Costs

Further Readings

Alreck, P., & Settle, R. B. (1995). *The survey research handbook: Guidelines and strategies for conducting a survey*. Chicago: Irwin.

Buckingham, A., & Saunders, P. (2004). *The survey methods workbook: From design to analysis*. Cambridge, UK: Polity Press.

Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. F., Martin, J., Nicholls, W. L., et al. (Eds.). (1998). *Computer assisted survey information collection*. New York: Wiley.

Research Question

A research question is an operationalization of the purpose(s) to which a survey project aims. The research question(s) should state the research problem in a way that allow(s) for appropriate research methods to be applied to gathering and analyzing information to help answer the question(s) (i.e., solve the problem[s]). Ultimately it is the research questions that guide the entire design of a survey, including the population that is to be sampled, how it is sampled, what sample size will be used, what mode is used to gather data, what measures will be included in the survey instrument, and what analyses will be conducted.

Typically the research question is posed in terms of *What*, *Why*, or *How*. For example, *What are the major reasons that citizens disapprove of the President? Why do African Americans, on average, have lower levels of social-economic status than do whites? How are people affected by their criminal victimization experiences in the short and longer term?*

Research questions are much more likely to be identified formally (explicitly) in academic research and much less likely in commercial research. Once the research question is clearly articulated, and after a review of past research on the topic has been conducted, research hypotheses should logically follow. In many instances, however, survey researchers, such as pollsters, need not identify any hypotheses, and their research questions are not formalized and merely left implicit.

Paul J. Lavrakas

See also Research Design; Research Hypothesis

Further Readings

Babbie, E. (2006). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth.

Blaikie, N. (2000). *Designing social research: The logic of anticipation*. Cambridge, UK: Polity.

Residence Rules

A design objective in many surveys is to measure characteristics of the population and housing in specified geographic areas. Some surveys, in addition, attempt to estimate counts of the population in these areas. A key component of the process needed to accomplish these goals is a clear set of *residence rules*, the rules that determine who should be included and who should be excluded from consideration as a member of a household.

It is important to recognize that the issue of defining residence applies to individuals rather than to households. In a given household, the residence of some members may be clear, whereas for others the rules discussed in this entry may need to be applied.

The question of "who belongs here" is self-evident in a large majority of households. The people residing there have no other place of residence, and there is no one missing who might be included in the household. People for whom determining residence is not so simple include college students, "snowbirds," commuter workers, live-in household help, military personnel, and migrant workers.

Usual Residence

The U.S. Bureau of the Census has traditionally used the concept of "usual residence." This is defined as the place where the person lives and sleeps most of the time and may be different from his or her legal address or voting address. As a general principle, people should be counted (i.e., included in a survey) at their usual place of residence. Some people have no usual place of residence; these people should be counted where they are staying at the time of the survey.

This gives rise to a set of rules that can be used to address most situations. They include the following:

- *People away on vacation and business* should be counted at their permanent residence.
- *Students*
  - Boarding school students should be counted at their parental homes. (This is the major exception to the usual rule.)
  - Students living away at college should be counted where they are living at college and therefore excluded from their parental homes.
  - Students living at home while attending college are counted at home.
- *Nonfamily members* in the home are included if this is where they live and sleep most of the time. Included here are live-in household help, foster children, roomers, and housemates or roommates.
- *Military personnel* are counted where they live or sleep most of the time, which generally means that they are not included in their permanent residence households. For example, personnel serving in Iraq, while absent, are not included in the household where they resided before they left and to which they will return.
- *Hospitals, prisons, and other institutions*
  - Persons staying temporarily at a general hospital, including newborn babies, are included in the household population, thus at their permanent residence.
  - Persons in chronic or long-term disease hospitals, such as a tuberculosis sanitarium or a mental hospital, are counted as living at the hospital, not at their previous permanent residence.
  - People in nursing or assisted-care homes are counted as living at the institution, not in the household from which they moved.
  - Inmates of prisons and jails, at any level, are counted at the penal institution and therefore not in their previous permanent residence household.

Some of the most difficult problems arise in the situations where a given person has more than one place of residence. This can happen because the person

- Has two (or more) homes,
- Has a permanent residence and another where he or she lives in order to commute to a job,
- Has a permanent residence and another where he or she lives in order to attend school, or
- Is a child living in joint custody, spending part of the time with one parent and the rest of the time with the other parent.

The general principle still applies in these situations: Where does the person live and sleep *most of the time*? This usually leads to counting the person at school (except for boarding school students), at the place where he or she lives to commute to the job, at the permanent residence rather than the vacation residence, and with the parent with whom the child spends the most time. When the question appears unanswerable, the general rule is to include the person where he or she is staying at the time of the interview.

These rules are derived from the standards set by the U.S. Bureau of the Census. However, there is a major case in which the bureau deviates from them: the American Community Survey (ACS). A 2-month residence rule applies for this survey. Its biggest effect is in situations where people have two residences and live in each part of the year. The ACS is a monthly survey conducted year-round. If the ACS sample finds the household in its second residence (e.g., in the state of Florida instead of the state of New York), and the household's total planned residence in the second home is at least 2 months, the residents are enumerated there. In parallel, if the sample hits the same household's permanent residence, it would be considered a "temporary vacancy." In the same situation in the decennial census, the household would be enumerated at its permanent residence, and the vacation home would be classified as a "vacant unit."

A place of residence is usually a *housing unit*, occupied by a *household* comprised of one or more persons, or vacant. It should be noted that not everyone lives in a household. Under current terminology, people who are not living in households are considered to be living in *group quarters* (unless they are literally homeless). Group quarters fall into two broad categories. *Institutional* facilities include prisons, nursing homes, mental hospitals, and other places where the residents are generally not free to come and go at will. *Noninstitutional* facilities include college dormitories, military barracks, and other places where the residents are free to move in, out, and about.

De Facto and De Jure Residence

The concepts of de facto and de jure residence are often used. The rules outlined here describe *de facto* residence, that is, the place where the person actually lives. *De jure* residence refers to the place where a person legally resides, for such purposes as estate taxes. The following are some classic cases where the two differ:

- College students usually maintain a de jure residence at their parental homes, which is the address used on their driver licenses and, often, on their voter's registration. In fact, the widespread implementation of motor–voter legislation has tied the two together.
- Senior citizens who spend most of their time in a warm climate, or even outside the United States, may maintain a de jure address for voting and estate purposes.
- Citizens living outside the United States often have a de jure voting address, from which they vote an absentee ballot.

In practice, survey research activities are rarely concerned with de jure residence. De facto residence, the place where people and households actually reside, is the relevant issue.

Patricia C. Becker

See also American Community Survey (ACS); U.S. Bureau of the Census

Further Readings

Panel on Residence Rules in the Decennial Census, Cork, D. L., & Voss, P. R. (Eds.), & National Research Council. (2006). *Once, only once, and in the right place: Residence rules in the decennial census*. Washington, DC: National Academies Press.

U.S. Census Bureau. (2003). *Plans and rules for taking the census: Residence rules*. Retrieved March 4, 2008, from http://www.census.gov/population/www/censusdata/resid_rules.html#resrules

Respondent

A respondent is the person who is sampled to provide the data that are being gathered in a survey. (In some social science disciplines, such as psychology, the person from whom data are gathered is called the "subject.") A respondent can report data about himself or herself or can serve as a proxy in reporting data about others (e.g., other members of the household). Even when the respondent is serving as a proxy, he or she is directly generating the data and contributing to their accuracy, or lack thereof.

Some surveys sample respondents directly, whereas other surveys begin by sampling larger units, such as households, and then choose a respondent within the unit from whom to gather data. In interviewer-administered surveying, such as what is done face-to-face or via the telephone, rapport first must be developed by the interviewer with the respondent in order to gain cooperation and then gather accurate data. In self-administered surveys, such as those conducted via mail and the Internet, there is no one representing the researcher's interests who is available to mediate the respondent's behavior, and cooperation typically is gained from the respondent via printed materials, such as a cover letter, which are sent to the respondent to read.

Gaining cooperation from sampled respondents has become progressively more difficult in the past two decades as lifestyles have become more hectic. The quality of the data that a survey gathers will be no better than the quality of the effort the respondent makes and her or his ability to provide it. A respondent's willingness and ability to provide accurate data will vary considerably across respondents and also from time to time for the same respondent.

There are many factors that influence whether a respondent will agree to participate in a survey and whether a respondent will provide complete and accurate data after agreeing to participate. It remains the responsibility of the researcher to choose the best survey methods, within the limitations of the researcher's finite budget for conducting the survey, that make it most likely that a sampled respondent will agree to cooperate when contacted, and once the respondent agrees, that he or she will provide the highest possible quality of data when answering the questionnaire.

Paul J. Lavrakas

See also Cover Letter; Proxy Respondent; Respondent–Interviewer Rapport; Respondent-Related Error; Within-Unit Selection

RESPONDENT BURDEN

The degree to which a survey respondent perceives participation in a survey research project as difficult, time consuming, or emotionally stressful is known as respondent burden. Interview length, cognitive complexity of the task, required respondent effort, frequency of being interviewed, and the stress of psychologically invasive questions all can contribute to respondent burden in survey research. The researcher must consider the effects of respondent burden prior to administering a survey instrument, as too great an average burden will yield lower-quality data and is thereby counterproductive.

Mechanisms that researchers may use to minimize respondent burden include pretesting, time testing, cognitive interviewing, and provision of an incentive. With pretesting, cognitive interviewing, and debriefing of respondents (and sometimes of interviewers as well) after the completion of the pretest, a researcher may glean insights into how to reduce any especially onerous aspects of the survey task. For example, it may be possible to break up a long series of questions that uses the same response format into two or three shorter series and space them throughout the questionnaire so that they do not appear and are not experienced as overly repetitive and, thus, as burdensome to complete. With sensitive and otherwise emotionally stressful questions, special transition statements that an interviewer relates to the respondent prior to these being asked may lessen the burden. Using incentives, especially contingent (i.e., performance-based) incentives, as well as noncontingent ones, often will raise the quality of the data that are gathered when the survey task is burdensome, as the respondents will strive to reach the level of data quality sought by the researcher, either because it will qualify him or her for an incentive or because he or she has a heightened feeling of obligation to the researcher to provide good quality data, or both.

Reduction of respondent burden may result in decreased nonresponse both at the unit level and the item level. When incentives are offered, data quality also should increase, as fewer respondents will turn to satisficing strategies to get them through the survey task as easily as possible regardless of how well they are complying with the task.

Ingrid Graf

See also Cognitive Interviewing; Debriefing; Incentives; Pilot Test; Respondent Fatigue; Response Rates; Satisficing

Further Readings

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

RESPONDENT DEBRIEFING

Respondent debriefing is a procedure that sometimes is carried out at the end of a survey's data collection phase. That is, when an individual participant has completed all aspects of the survey, debriefing occurs for that person. Debriefing is usually provided in the same format as the survey itself, that is, paper and pencil, online, or verbally via telephone. There are two major reasons that a researcher may want to debrief respondents: (1) The researcher may want to gather feedback from the respondent about the respondent's experience participating in the study or about more details concerning the topic of the survey, and (2) the researcher may have used some form of deception as part of the study and will use the debriefing to inform the respondent of this and try to undo any harm the deception may have caused the respondent.

Typically, the debriefing begins with a statement of gratitude for the participation of the respondent. This is followed by a brief restatement of the objective of the survey that would have initially been provided to the participant at the time that informed consent was being sought. Third, an overview of the tasks completed during the survey is given. Fourth, a very general statement of what will be done with all participants' responses is provided. Fifth, the participant is given the chance to request a report of the results of the survey when data analysis is complete, and contact information must be collected so that the results of the survey can be provided the participant.

In addition, appropriate referrals relative to the nature of the survey should be provided. For example, in research into depression among college students, some researchers provide appropriate referrals, including the campus counseling center and the local county mental health agency. Finally, the name of the principal investigator is provided, as well as his or her contact information, in the event that the respondent would need to communicate a concern or ask questions about the survey; contact information should include office phone number, work email address, and office address.

After data analysis has been completed, either a report of the survey results is mailed to survey participants or the results of the survey are published on a Web site to which survey participants are provided the Web address, perhaps via postcard or email.

In the event that the research involved the use of deception, the debriefing phase must also attempt to undo the effect of the deception by acknowledging that false information was given to some survey participants. Those who were deceived must be told that fact, regardless of how slight the harm may be in the eyes of the researcher.

Carla R. Scanlan

See also Debriefing; Deception; Informed Consent

Further Readings

Campanelli, P. C., Martin, E. A., & Rothgeb, J. M. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. *The Statistician*, *40*(3), 253–264.

Hughes, K. A. (2003, November). *Comparing pretesting methods: Cognitive interviews, respondent debriefing and behavioral coding*. Paper presented at the annual meeting of the Federal Committee in Statistical Methodology, Arlington, VA. Retrieved May 18, 2008, from http://0-www.census.gov.mill1.sjlibrary.org/srd/papers/pdf/rsm2004-02.pdf

Toy, D., Wright, L., & Olsen, L. (2001). A conceptual framework for analyzing deception and debriefing effects in market research. *Psychology and Marketing*, *18*(7), 691–719.

Respondent-Driven Sampling (RDS)

Respondent-driven sampling (RDS) is a method for drawing probability samples of "hidden," or alternatively, hard-to-reach, populations. Populations such as these are difficult to sample using standard survey research methods for two reasons: First, they lack a sampling frame, that is, an exhaustive list of population members from which the sample can be drawn. Second, constructing a sampling frame is not feasible because one or more of the following are true: (a) The population is such a small part of the general population that locating them through a general population survey would be prohibitively costly; (b) because the population has social networks that are difficult for outsiders to penetrate, access to the population requires personal contacts; and (c) membership in the population is stigmatized, so gaining access requires establishing trust. Populations with these characteristics are important to many research areas, including arts and culture (e.g., jazz musicians and aging artists), public policy (e.g., immigrants and the homeless), and public health (e.g., drug users and commercial sex workers).

These populations have sometimes been studied using institutional or location-based sampling, but such studies are limited by the incomplete sampling frame; for example, in New York City only 22% of jazz musicians are musician union members and they are on average 10 years older, with nearly double the income, of nonmembers who are not on any public list.

This entry examines the sampling method that RDS employs, provides insights gained from the mathematical model on which it is based, and describes the types of analyses in which RDS can be used.

Sampling Method

RDS accesses members of hidden populations through their social networks, employing a variant of a snowball (i.e., chain-referral) sampling. As in all such samples, the study begins with a set of initial respondents who serve as "seeds." These then recruit their acquaintances, friends, or relatives who qualify for inclusion in the study to form the first "wave." The first wave respondents then recruit the second wave, who in turn recruit the third wave, and so forth. The sample expands in this manner, growing wave by wave, in the manner of a snowball increasing in size as it rolls down a hill.

RDS then combines snowball sampling—a nonprobability sampling technique—with a mathematical model that weights the sample to compensate for the fact that it was not obtained in a simple random way. This procedure includes controls for four biases that are inherent in any snowball sample:

1. The seeds cannot be recruited randomly, because if that were possible, the population would not qualify as hidden in the first place. Generally, the seeds are respondents to whom researchers have easy access, a group that may not be representative of the full target population. Consequently, the seeds introduce an initial bias.
2. Respondents recruit their acquaintances, friends, and family members, whom they tend to resemble in income, education, race/ethnicity, religion, and other factors. This homophily principle was recognized by Francis Galton more than a century ago. Its implication is that by recruiting those whom they know, respondents do not recruit randomly. Instead, recruitments are shaped by the social network connecting the target population. Consequently, successive waves of recruitment introduce further bias into the sample.
3. Respondents who are well connected tend to be oversampled, because more recruitment paths lead to them. Therefore, higher-status respondents—those who have larger social networks—are oversampled.
4. Population subgroups vary in how effectively they can recruit, so the sample reflects, disproportionately, the recruitment patterns of the most effective recruiters. For example, in AIDS prevention research, HIV positives generally recruit more effectively and also tend to recruit other positives, so positives tend to be oversampled.

Mathematical Model

RDS is based on a mathematical model of the network-recruitment process, which functions somewhat like a corrective lens, controlling for the distorting effects of network structure on the sampling process to produce an unbiased estimate of population characteristics. Space here does not permit presentation of the mathematical model on which RDS is based, but two insights upon which it is based provide a sense for how the model operates. First, modeling the recruitment process as a regular Markov chain reveals that if referral chains are sufficiently long, that is, if the chain-referral process consists of enough waves, the composition of the final sample becomes independent of the seeds from which it began. The point at which the sample composition becomes stable is termed the *equilibrium*. Therefore, an important design element in RDS involves measures for increasing the length of referral chains. Means for creating long chains include that respondents be recruited by their peers rather than by researchers, providing rewards for peer recruiters, and setting recruitment quotas so a few cannot do all the recruiting. Through these means, a major concern is resolved regarding bias in chain-referral samples, that is, producing a population estimate that is independent of the seeds (initial subjects) with which the sampling began.

Second, gathering information about the network structures through which the sampling process expands provides the means for controlling for the biasing effects of those structures. This procedure entails a weighing process that quantifies the biasing effects of network structure, to compensate for oversampling groups with higher levels of homophily (i.e., network segmentation), groups having larger social networks, and those favored by the recruitment patterns of the most effective recruiting groups. These three potential sources of bias can operate in the same or opposite directions, and the model calculates the balance among them.

If the assumptions upon which RDS is based are satisfied, the sample is asymptotically unbiased, which means that bias is on the order of 1/(sample size), so bias is trivial in samples of meaningful size. The model is based on five assumptions. The first three specify the conditions under which RDS is an appropriate sampling method. First, respondents must know one another as members of the target population. Peer recruitment is a feasible sampling strategy only if this condition is satisfied. Consequently, RDS would not be suitable for sampling tax cheats, who can be friends and not know they share membership in that hidden population. On the other hand, it is suitable for sampling populations linked by a "contact pattern," such as musicians who perform together or drug users who purchase drugs together. Second, ties must be dense enough to sustain the chain-referral process. For populations linked by a contact pattern, this is rarely problematic, but it may be problematic for other groups. Third, sampling is assumed to occur with replacement, so recruitments do not deplete the set of respondents available for future recruitment. Consequently, the sampling fraction should be small enough for a sampling-with-replacement model to be appropriate.

The fourth assumption states that respondents can accurately report the number of relatives, friends, and acquaintances who are members of the target population. Studies of the reliability of network indicators suggest that this is one of the more reliable indicators; furthermore, the RDS population estimator depends not on absolute but on relative degree, so variations in name generators that inflate or deflate the reports in a linear manner have no effect on the estimates.

Finally, the fifth assumption specifies that respondents recruit as though they are choosing randomly from their networks. This is based on the expectation that respondents would lack an incentive or ability to coordinate to selectively recruit any particular group, and support for this expectation has been found. The plausibility of this assumption is determined, in part, by appropriate research design. For example, if a research site were located in a high-crime neighborhood, recruiting residents of the neighborhood might be easy, but recruiting peers from more comfortable neighborhoods might prove difficult, so sampling would be nonrandom because it excluded the latter group. However, if research identifies neutral turf in which all potential respondents feel safe, the random recruitment assumption is made more plausible. Similarly, if incentives are offered that are salient to respondents from all income groups (e.g., a choice between receiving a monetary reward and making a contribution to a charity of the respondent's choice), the random recruitment assumption is made more plausible.

Additional Analyses

RDS emerged as a member of a relatively new class of probability sampling methods termed *adaptive* or *link-tracing* designs, which show that network-based sampling methods can be probability sampling methods. RDS continues to evolve. Though originally limited to univariate analysis of nominal variables, it has been extended to permit analysis of continuous variables and multivariate analysis. The significance of these methods is that they expand the range of populations from which statistically valid samples can be drawn, including populations of great importance to public policy and public health.

Douglas D. Heckathorn

See also Adaptive Sampling; Network Sampling; Nonprobability Sampling; Probability Sample; Sampling Frame; Snowball Sampling

Further Readings

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, *44*(2), 174–199.

Heckathorn, D. D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, *37*, 151–207.

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, *35*, 193–238.

Thompson, S. K., & Frank, O. (2000). Model-based estimation with linktracing sampling designs. *Survey Methodology*, *26*(1), 87–98.

Volz, E., & Heckathorn, D. D. (2008). Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics*, *24*(1), 79–97.

Respondent Fatigue

Respondent fatigue is a well-documented phenomenon that occurs when survey participants become tired of the survey task and the quality of the data they provide begins to deteriorate. It occurs when survey participants' attention and motivation drop toward later sections of a questionnaire. Tired or bored respondents may more often answer "don't know," engage in "straight-line" responding (i.e., choosing answers down the same column on a page), give more perfunctory answers, or give up answering the questionnaire altogether. Thus, the causes for, and consequences of, respondent fatigue, and possible ways of measuring and controlling for it, should be taken into account when deciding on the length of the questionnaire, question ordering, survey design, and interviewer training.

Participating in a survey requires time and effort; respondents often need to reflect on their behaviors, retrieve or construct opinions on issues, and evaluate candidates, policies, or products. As the time to complete the survey grows longer, the motivation and ability needed by respondents to accurately answer the questions may decline. The level of processing required to answer the questions may also induce fatigue, such that as the questions are more detailed, require recalling past events, comparing or choosing between many different options, motivation may wear thin. Another factor that can generate fatigue is the specific topic of the survey: how interesting or important it is to participants and the type of interaction they have with the interviewer about it. Generally speaking, as (a) the survey is more time consuming, (b) the questions are boring and complicated, (c) more open-ended questions are asked, (d) the interviewer does not motivate adequate answers, and (e) the issue of the survey is mundane or repetitive, respondents' motivation may decrease and fatigue effects may arise.

Fatigue effects may have several consequences for the later items of a questionnaire. Respondents in self-administered surveys may fail to read adequately the lists of response alternatives, skip questions more frequently, or be more likely to engage in satisficing by choosing answers such as "not applicable" or "don't know." Fatigue may also cause more stereotypical answers, known as straight-line (or response set) responding; these occur when a series of consecutive questions share the same answer choices that appear in the same order, such that an unmotivated person may answer with the same response on all items in the series.

There are several ways in which researchers try to measure and assess fatigue effects. First, the questionnaire may be split and the order of the questions may be randomized or counterbalanced in the different versions. The responses to items presented late on one version can be then compared to responses to the same items when presented earlier, in terms of percentage of nonresponses, don't knows, straight-line responding, and correlations with other variables. Another option to evaluate whether fatigue effects took place is by measuring the consistency of responses to repeated questions appearing early and late in the questionnaire, that is, including an alternative wording for some of the questions and measuring their reliability with the questions appearing earlier.

Perhaps the simplest and best way of dealing with fatigue effects is to avoid them. Although some research suggests that later items in a lengthy questionnaire may be systematically more vulnerable to poor response rates and to inadequate answers, it seems the fatigue effect may be avoided as long as respondent's motivation is maintained. Thus, a researcher should try to balance between collecting accurate and sufficient information and conducting a well-structured and not-too-long questionnaire.

Pazit Ben-Nun

See also Don't Knows (DKs); Nonresponse Bias; Nonsampling Error; Respondent Burden; Respondent-Related Error; Response Order Effects; Satisficing

Further Readings

Bradburn, N. M., & Mason, W. M. (1964). The effect of question order on responses. *Journal of Marketing Research*, *1*, 57–64.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*, 549–559.

Respondent–Interviewer Rapport

Respondents and interviewers interact during the conduct of surveys, and this interaction, no matter how brief, is the basis for a social relationship between the two. Often this relationship begins when the interviewer calls or visits the respondent in an attempt to initiate and complete an interview. Other times, the respondent may call the interviewer in order to complete an interview. During the social interaction of conducting an interview, the respondent and interviewer will typically develop a rapport.

The establishment of rapport between the respondent and the interviewer, or lack thereof, is a key element in the interviewer gaining the respondent's cooperation to complete an interview. If a good rapport is not established, the likelihood of the interviewer completing an interview decreases. Further, good rapport will make the respondent comfortable with answering questions that could be considered personal or embarrassing. It is important for interviewers to convey a neutral, nonjudgmental attitude toward respondents regardless of the survey topic or the content of the respondents' answers in order to make the respondent as comfortable as possible. A respondent who is comfortable is more likely to provide accurate responses and take the survey request seriously.

The rapport between respondents and interviewers plays a large role in the interview process. Interviewers are largely initially responsible for developing and maintaining rapport with respondents. The respondent has to feel comfortable with the survey, and this can be influenced by how key information about what the survey requires is delivered by the interviewer. The respondent must also believe that the survey request is legitimate and the data they provide will be protected. Interviewers must explain the reason for the survey, set the tone for the interview, convey the importance of the survey, and set up the expectations for how the interview will proceed from the start of the interview. Further, a respondent's decision to participate, as well as his or her attitude about participating (e.g., the seriousness of the survey request) is impacted by his or her feelings about the interviewer and the rapport that has been established.

Rapport is usually established during the first few seconds of a call or visit; however, the rapport should continue to build throughout the interaction. Interviewers should be trained in ways to quickly establish a rapport with the respondents. Gaining cooperation to complete an interview through establishing rapport is the target outcome of each respondent–interviewer interaction, and this target outcome is the same for both in-person and telephone interviews. In both interview modes, it is important that interviewers convey a friendly, professional tone. Interviewers must also be sincere, confident, knowledgeable, and well prepared for their interactions with respondents. All of these interviewer characteristics taken together will impact the kind of rapport established with the respondent. However, the communication channels through which interviewers have to convey these key characteristics differ based on interview mode. For in-person interviews, respondents have all communication channels with which to establish rapport with the interviewer, including verbal and nonverbal. Conversely, for telephone interviews, the respondent has only verbal communication channels (which involves two things: the interviewers' voice characteristics and the words they say) on which to judge the interviewer. Because of these differences in communication channels, training for in-person and field interviewers differs somewhat on how to convey the key interviewing characteristics.

Beyond the social rapport–building skills of interviewers, there are also ways to write a questionnaire to increase the likelihood that the rapport established during the gaining cooperation phase will continue throughout the survey administration. For example, asking interesting questions early on in the interview can increase the respondents' interest in the survey. Asking questions that are relatively simple to answer can increase the respondents' comfort with their role in the task and can influence their willingness to continue with the interview. Conversely, placing difficult-to-answer or sensitive questions too early in the interview can discourage respondents from continuing the interview. Although these questionnaire design techniques can prove effective in building or strengthening respondent–interviewer rapport, they are not utilized by all researchers.

Although in-person interviewers and telephone interviewers must use different channels of communication with respondents, many of the same techniques are taught to both types of interviewers. For example, trainers often use role playing, group or paired practice, and quizzes as ways to educate interviewers on building rapport.

No one interviewing style has been identified as being the best in establishing rapport and building relationships with respondents. While interviewers are trained on basic techniques that have been shown to have success in establishing rapport with the respondent and setting up the expectations for the interview, interviewers are often instructed to use these techniques in a way that is comfortable for them. The most important factors in establishing and maintaining a relationship with respondents are clearly communicating the role expectations for the relationship and setting up standards for the conduct of the interview early on in the interaction.

Lisa Carley-Baxter

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Interviewer; Respondent

Further Readings

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Respondent Refusal

The respondent refusal disposition is used in telephone, in-person, mail, and Internet surveys to categorize a case in which contact has been made with the designated respondent, but he or she has refused a request by an interviewer to complete an interview (telephone or in-person survey) or a request to complete and return a questionnaire (mail or Internet survey). A case can be considered a respondent refusal only if the designated respondent has been selected and it is clear that he or she has stated he or she will not complete the interview or questionnaire. Respondent refusals are considered eligible cases in calculating response and cooperation rates.

In a telephone survey, a case is coded with the respondent refusal disposition when an interviewer dials a telephone number, reaches a person, begins his or her introductory script, and selects the designated respondent, and the respondent declines to complete the interview. In calls ending in a respondent refusal, the designated respondent may provide an explanation for the refusal such as, "I don't do surveys," "I don't have time," "I'm not interested," or "Please take me off your list." In other instances, the respondent contacted may simply hang up.

Respondent refusals in an in-person survey occur when an interviewer contacts a household, a household member answers the door, the interviewer begins his or her introductory script, and selects the designated respondent, and the designated respondent declines to complete the interview. Common explanations in in-person surveys for household refusals parallel those for telephone surveys.

Cases in a mail or Internet survey of specifically named persons are coded with the respondent refusal disposition when contact has been made with the sampled person and he or she declines to complete and return the questionnaire. Because little may be known in a mail survey about who in the household generated the refusal, it can be difficult to determine whether a household refusal or respondent refusal disposition is most appropriate. Different invitation methods for Internet surveys (such as contacting sampled respondents at their email addresses) make respondent refusals the most common type of refusal in an Internet survey.

Respondent refusals usually are considered a final disposition. Because refusal rates for all types of surveys have increased significantly in the past decade, many survey organizations review cases ending in respondent refusals and select cases in which the refusal is not extremely strong in nature to be contacted again in order to try to convert the case's disposition to a completed interview.

Matthew Courser

See also Final Dispositions; Household Refusal; Refusal Conversion; Refusal Report Form (RRF); Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research.* Chicago: University of Chicago Press.

Respondent-Related Error

Respondent-related error refers to error in a survey measure that is directly or indirectly attributable to the behaviors or characteristics of respondents; it is distinguished from error resulting from other survey components, such as questionnaires, interviewers, or modes of administration. However, respondents may interact with these other components of survey design in producing errors. It is useful to dichotomize respondent-related errors into those that arise from nonobservation or nonresponse (e.g., during efforts to obtain interviews) and those that result from observation or measurement (e.g., during the administration of the survey).

Errors in survey measures have two components: bias and variable errors. Bias results when responses provided in the survey differ from their true values, which are typically unknown and unmeasured, in a systematic way across repeated measurements. For example, in contrast to the responses they enter onto self-administered questionnaires, respondents are more likely to underreport abortions they have had when data are collected by an interviewer. Variable errors result from differences across the source of the error.

For example, respondents sometimes provide different answers to interviewers who deviate from the rules of standardized interviewing. Survey methodologists also make a distinction between error associated with objective versus subjective questions. For objective questions about events and behaviors, error is expressed as the difference between the respondent's answer and what might have been observed if observation had been possible. For subjective questions about attitudes and opinions, error is conceptualized as sources of variation in the answers other than the concept the researcher is trying to measure.

Respondents and Nonresponse Error

Respondents may contribute to nonresponse bias when they refuse to participate in the study or cannot be located or contacted. Nonresponse bias varies as a function of the survey's response rate and the degree to which nonresponders differ from participants. Obtaining high response rates is generally considered an important protection from nonresponse bias. Nonetheless, nonresponse bias may be large, even with a high response rate, if those interviewed differ substantially from those who are sampled but are never contacted or those who refuse; conversely, bias may be small, even with a low response rate, if respondents are similar to noncontacts and refusers on the characteristics of interest. In longitudinal studies, nonresponse error also varies if respondents who drop out of the study systematically differ from respondents who are retained in the panel (i.e., so-called differential attrition).

Attempts to estimate the impact of different levels of response on survey estimates and nonresponse error suggest that improvements in response rates sometimes have a negligible impact on both estimates and error, but the results are unpredictable. At present there is little empirical or theoretical guidance to predict how much nonresponse or under what circumstances nonresponse will produce nonresponse error, but nonresponse can have a big impact on survey error, especially for some subgroups.

Even after they agree to participate in a survey, respondents may fail to provide data within the interview; they may do this intentionally, by refusing to answer questions or saying "don't know," or unintentionally, by skipping questions on self-administered questionnaires. Item nonresponse occurs more often for some types of questions than others. For example, refusals are more common for questions about income and for questions that ask for sensitive or threatening information. Respondents may also provide responses that are incomplete, for instance, by not providing enough information to an open-ended question to allow their answers to be classified reliably.

Respondents and Measurement Error

Respondents contribute to measurement error in several ways, including the characteristics they bring to the interview, such as their cognitive skills or motivation; their behavior within the interview; and their interaction with the survey instrument and with interviewers. Certain demographic characteristics may be associated with errors. For example, some research indicates that older and male respondents may be more likely to misreport than younger and female respondents, respectively. Respondents' characteristics may interact with interviewers' characteristics and the topic of the survey to produce response effects. An oft-cited finding is that respondents report differently to interviewers of the same race or gender when the topic concerns attitudes about race or gender than they do to interviewers of a different race or gender. A respondent's level of motivation may be particularly important for questions about events and behaviors that require searching memory. Retrospective questions require respondents to remember information that may be difficult to retrieve or emotionally charged. Fatigue and boredom may also affect the data quality of survey reports and thereby lead the respondent to providing merely an "easy" answer (i.e., satisficing).

Roger Tourangeau and his colleagues have described the question–answer process using a model to describe the stages respondents go through when answering a survey question. The first stage, *encoding*, occurs before the survey when respondents store information about an event or experience into memory. The effectiveness of the initial encoding can have great bearing on the accuracy of a respondent's survey report. Errors due to encoding failures occur when respondents are asked to report about information they may not have encoded at all or only partially encoded, such as information about their children's immunizations, or when proxy respondents are used to report about others.

In the second stage of the response process, *comprehension*, respondents either hear or read the question and attempt to understand its words and phrases, general meaning, and the intent of the question writer. Respondent-related errors occur frequently at this stage. Experiments as well as research from cognitive interviewing demonstrate that the meaning of concepts within a question is frequently understood in unintended ways. For instance, context effects occur when respondents use information provided by previous questions or answers in interpreting meaning and formulating a response for the current question. Other errors at the comprehension stage can result if (a) questions contain vague terms or phrases that are too technical, (b) respondents lack an opinion on the topic, or (c) the questions are not understood as the researchers intended because the questions were poorly translated or because the respondent has vision or hearing problems. Respondents in interviewer-administered surveys may attempt to elicit the interviewer's help in resolving a comprehension problem. Traditionally, interviewers are trained to treat requests for clarification with the probe "whatever it means to you," which may not solve the comprehension problem. If, however, the interviewer responds to the respondent's request in a way that departs from the rules of standardized interviewing, the interviewer–respondent interaction may be another source of error.

Retrieval, the third stage in the question–answer process, involves accessing information from an inactive state in long-term memory and bringing it into working memory, a state of activation in which the information can be used. For objective questions that ask about events and behaviors, retrieval may involve recalling the number of times the respondent engaged in a particular action during a given time frame. For subjective questions, retrieval may involve recalling previously formulated opinions, attitudes, or feelings. For some types of questions, retrieval may also include consultation with external records, such as for reporting income, or consultation with others, such as for proxy reports.

Retrieval can influence response accuracy. For example, respondents engage in forward telescoping when they remember events or behaviors as having happened earlier than they actually took place; in backward telescoping, events are remembered as occurring later than they did. Errors due to omission also occur frequently, even for highly salient events such as hospitalizations. Events are best remembered when they are salient (i.e., unusual, costly, or enduring), infrequent, rare, and dissimilar from other events.

Several factors appear to be influential in increasing the effort expended by respondents during retrieval and in improving the accessibility of information stored in memory. Evidence suggests that the more time respondents are given to respond, the better their responses will be. This is probably attributable to the fact that recalling and counting events is an effortful process that can be discouraged both by the pace of the interview and by respondents themselves, who may want to finish quickly. Other methods that can increase respondent effort and accuracy are soliciting formal respondent commitment, limiting time frames to periods a respondent can reasonably access from memory, giving respondents multiple opportunities to answer a question (like in a "second guess"), and using longer questions that provide more time for respondents to process and retrieve answers. As an example, longer questions have been associated with better reporting for health events and higher reports of threatening behaviors like drinking and sexual activity.

In the fourth stage of the response process, *judgment* and *mapping*, respondents evaluate the information retrieved from memory and formulate a candidate response vis-à-vis the format of the survey question. Research suggests that the format of a question, such as whether it is open-ended or closed-ended, can influence respondent-related errors. Most survey questions are closed-ended and offer a set of predetermined response categories. A question on satisfaction might include the categories, "extremely satisfied/dissatisfied," "satisfied/dissatisfied," and "somewhat satisfied/dissatisfied." In assessing behavioral frequencies, however, closed-ended questions that offer response categories are associated with several undesirable outcomes. Categories may indicate to respondents the range of "acceptable" responses. The wording of response categories can decrease effort by indicating that an exact answer is not sought. The number and wording of verbal categories may decrease accessibility of memories by placing too great a load on cognitive processing, especially in a telephone survey where respondents cannot read and reread answer categories to ensure more thorough processing.

Many other kinds of response effects occur during retrieval, and these may interact with the mode of administration. For example, primacy effects, the tendency to select options at the beginning of a set of categories, are found in face-to-face interviews where questions with several options are often presented visually on show cards. In contrast, recency effects, the tendency to select options at the end of the scale, occur in telephone interviews in which questions are presented orally.

In the last stage of the response process, respondents report their answer. Respondents demonstrate many characteristic ways of "reporting" that are biasing. Some of the most common are providing answers that are socially desirable, satisficing, acquiescing, and intentionally misreporting in order to protect privacy (e.g., not reporting income for fear the results will be shared with the Internal Revenue Service). Social desirability bias refers to error resulting from respondents' tendencies to overreport attitudes or behaviors deemed socially desirable, such as voting, and underreport those judged undesirable, such as illegal drug use. Respondents may misreport for several reasons, possibly because they are used to doing so, or in order to present themselves positively to an interviewer or the researcher. Misreporting of answers has been documented for many diverse topics, including abortion, drug and alcohol use, sexual behaviors, voting and attention to politics, as well as attitudes about sensitive issues such as race. Many studies have explored the effect of data collection mode on reporting threatening questions. In comparison to reporting in self-administered interviews, respondents in interviewer-administered interviews are more likely to provide lower reports of these behaviors. Social desirability bias can be reduced and reporting accuracy can be enhanced by increasing respondents' trust in the legitimacy of the organization sponsoring the survey, by raising respondents' perceptions of the importance of the answers they are providing, and by heightening respondents' sense of privacy and confidentiality. For example, in face-to-face surveys, interviewers can provide respondents with a computer into which they can enter their responses directly, as is done with computerized self-administered questionnaires, rather than having respondents report sensitive or threatening information directly to the interviewer.

Acquiescing is the tendency of respondents to agree to or passively accept a proposition offered by the question. Satisficing is similar but somewhat broader. Satisficing occurs when respondents engage in the minimum amount of processing necessary to respond to a question, but without wholly investing in providing the most accurate answer possible. It is manifested when respondents choose an option such as "don't know" or when respondents repeatedly select the same or similar scale points in a battery of questions or in a scale. Both acquiescing and satisficing can reflect a lack of respondent motivation or ability or the difficulty of the survey task.

Methods for Studying Respondent-Related Error

What survey researchers know about respondent-related error comes from several paradigms for studying response errors. In record-check studies, survey reports are compared to external records (such as court records in a study of child support) to assess how accurately respondents report. Split-half experiments are conducted during interviews to test different wordings of questions and response categories. Qualitative methods such as focus groups, in-depth interviews, and cognitive interviews are conducted to determine how well respondents' understanding of survey concepts matches that of the question writers. Cognitive interviews are also used to understand retrieval and judgment processes. For interviewer-administered surveys, the interaction between the respondent and interviewer is recorded and analyzed. The methods for studying response errors are costly and complex and sometimes do not suggest a clear method for reducing error.

Jennifer Dykema, Steven Blixt, and John Stevenson

See also Acquiescence Response Bias; Closed-Ended Questions; Cognitive Interviewing; Comprehension; Computerized Self-Administered Questionnaires (CSAQ); Context Effect; Differential Attrition; Encoding; Measurement Error; Nonresponse Error; Open-Ended Question; Primacy Effect; Proxy Respondent; Recency Effect; Record Check; Respondent Burden; Respondent Fatigue; Response Bias; Retrieval; Satisficing; Social Desirability; Split-Half; Telescoping

Further Readings

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (1991). *Measurement errors in surveys*. New York: Wiley.

Biemer, P. P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Response

In survey research, a response generally refers to the answer a respondent provides when asked a question. However, a response to a survey can also be considered to occur when a sampled respondent decides whether or not to participate in the survey. Both types of responses may affect data quality; thus, a well-trained researcher will pay close attention to each.

Response to the Request to Participate

At the time a person is notified of having been sampled to participate in a survey, he or she may decide to participate or not. This response, whether or not to participate, affects the response rate of the survey, which essentially is the proportion of eligible respondents who agree to participate. Response rates may vary as the result of a number of factors, including the population being studied, the procedures used for sampling, the mode of the survey (e.g., telephone, in person, mail, or Web), questionnaire length, field period length, number of contact attempts, the topic of the survey, and whether procedures such as advance letters, refusal conversions, and incentives are used.

Although no specific response rates are generally required for conducting a survey, researchers traditionally strive for high response rates. Low response rates may lead to nonresponse bias when certain groups of respondents are more likely to participate than others, which creates the phenomenon of under- or overrepresentation of the certain attributes that the survey is striving to measure. As such, it is in a researcher's interest to encourage an affirmative response to the request to participate from as many respondents as possible that are sampled.

Response to Survey Questions

When someone is asked a survey question, the effort that goes into determining and then providing an answer (the response) can vary considerably from respondent to respondent. Thus, respondent-related error is of considerable consequence to survey researchers. Such error may be introduced into survey responses to individual survey questions in a number of ways, such as when respondents answer survey questions based on what the survey designers expect. This may occur because respondents want to please researchers or because of leading questions. Response error may also be introduced if respondents misinterpret or do not understand a question. The format of survey questions (e.g., whether they are open- or closed-ended, the number and specific response choices provided) influences respondents' comprehension of a survey question. Response error can be minimized if researchers consciously implement strategies while designing the instrument, such as avoiding leading questions and wording questions precisely. The order of questions in surveys can also contribute to response error, although there are also some general guidelines for minimizing error from this source (e.g., easy questions should be placed before hard ones).

Respondents themselves can also be a source of response error because respondents may not be willing or able to answer the questions with correct information. Thus, response error may be introduced if respondents satisfice or engage in socially desirable responding. These issues can also be somewhat alleviated through optimal questionnaire design.

Cary Stacy Smith and Li-Ching Hung

See also Questionnaire Design; Question Order Effects; Respondent; Respondent-Related Error; Response Rates; Satisficing; Social Desirability

Further Readings

de Leeuw, E. D. (1992). *Data quality in mail, telephone, and face to face surveys*. Amsterdam: TT Publications.
Fink, A. (2002). *The survey handbook* (2nd ed.). Thousand Oaks, CA: Sage.

Response Alternatives

Response alternatives are the choices that are provided to respondents in a survey when they are asked a question. Response alternatives are generally associated with closed-ended items, although open-ended items may provide a limited number of such choices.

Response alternatives can take a number of different forms, related to the type of question presented. In a Likert-type item, in which respondents are asked the extent to which they agree or disagree with a statement, the response alternatives might be s*trongly approve, approve, neither approve nor disapprove,*

disapprove, and *strongly disapprove*. For a rating scale—for example, the rating of the job the president is doing—the response alternatives might be *excellent, good, fair, poor*, and *very poor*. Another example of a rating scale would be a "feeling thermometer" in which respondents would be asked to rate some individual or object on a scale from 0° to 100°, in which the degrees would represent the response categories.

Types of response alternatives may vary depending on the method used for data collection. In face-to-face and self-administered surveys, respondents can be presented with visual materials displaying the response alternatives. For example, in a face-to-face survey, a respondent can be presented a show card describing the various alternatives, and a similar description of the response options can be provided in a mail or other self-administered survey and for surveys conducted via the Internet. As a result, a larger number of response alternatives can be presented in these modes of data collection than in telephone surveys. The number of response alternatives that an average respondent can remember over the telephone is generally limited to five. If a larger number of response alternatives is desired, the question is typically divided into a root item, which is then "unfolded" into this larger number of choices. For example, if researchers were interested in a person's ideology along a liberal–conservative scale, they might first ask, *In politics, do you generally consider yourself liberal, moderate, or conservative?* Those who said "liberal" would then be asked, *Would you say you are extremely liberal or somewhat liberal?* and, similarly, those who said "conservative" would be asked, *Would you say you are extremely conservative or somewhat conservative?* Respondents who answered "moderate" would be asked, *Do you lean toward the liberal side or lean toward the conservative side?* The result would be seven response alternatives: (1) extremely liberal, (2) somewhat liberal, (3) leans toward liberal, (4) moderate—leans toward neither, (5) leans toward conservative, (6) somewhat conservative, and (7) extremely conservative.

Another consideration in presenting response alternatives is whether to provide choices such as "don't know" and "refused" as explicit options to respondents. Typically such choices are not read to respondents in face-to-face or telephone interviews and are not included as options in self-administered or Internet surveys. There are situations, however, in which "don't know" or "no opinion" responses are important to the researcher; in such cases an explicit "no opinion" response would be provided. For example, in the question, *Do you think state spending on roads and highways should be increased, kept about the same as it is now, decreased, or don't you have an opinion on this issue?* the response alternative *or don't you have an opinion on this issue* is an explicit "don't know" or "no opinion" option. When such options are provided as part of the question, the percentage of respondents who choose this alternative is higher than when it is not present, and respondents have to volunteer that they "don't know."

The fact that a larger percentage of respondents will choose a "don't know" response when offered explicitly applies more generally to all types of responses. For example, if survey respondents are presented with the question, *Do you think that federal government spending on national defense should be increased or decreased?* some percentage will volunteer that they believe it should be kept about the same as it is now. If the same group of respondents were asked the question, *Do you think federal government spending on national defense should be increased, decreased, or kept about the same as it is now?* a much higher percentage would select the "same as it is now" response alternative. Similarly, for items in which respondents are presented with some choices (generally the most likely options) but not an exhaustive list, researchers will often leave an "other" option. An example of such an "other" response is provide by the question, *Which television network do you watch most frequently, ABC, CBS, Fox, NBC, or some other network?* Although some respondents will provide other responses such as UPN or the CW, the percentage of such responses will typically be smaller than if these other networks had been mentioned specifically in the question.

An additional aspect of response alternatives is balance. In the design of most questions, researchers strive to provide balance in the response alternatives. If presenting a forced choice item, the strength of the arguments on one side of a question should be similar to the strength of the arguments on the other. Similarly, in asking an agree–disagree item, an interviewer would ask, *Do you agree or disagree that . . .*; asking only *Do you agree that . . .* would lead more respondents to choose this option. In asking about approval or disapproval of some individual or proposal, a question that asks, *Do you completely approve, approve a great deal, approve somewhat, or disapprove of . . .* is not balanced. Asking *Do you approve a great deal,* *approve somewhat, disapprove somewhat, or disapprove a great deal of*… provides a more balanced item.

Response alternatives are an important consideration in question design. The number of choices presented, how they are balanced, whether there is an explicit middle alternative, and an explicit "no opinion" option can all influence the results obtained by a survey question.

Robert W. Oldendick

See also Balanced Question; Closed-Ended Question; Feeling Thermometer; Forced Choice; Likert Scale; Open-Ended Question; Ranking; Rating; Response Order Effects; Show Card; Unfolding Question

Further Readings

Converse, J., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Beverly Hills, CA: Sage.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys*. Thousand Oaks, CA: Sage.

Selltiz, C., Jahoda, M., Deutsch, M., & Cook, S. (1959). *Research methods in social relations*. New York: Holt, Rinehart & Winston.

Response Bias

Response bias is a general term that refers to conditions or factors that take place during the process of responding to surveys, affecting the way responses are provided. Such circumstances lead to a nonrandom deviation of the answers from their true value. Because this deviation takes on average the same direction among respondents, it creates a systematic error of the measure, or bias. The effect is analogous to that of collecting height data with a ruler that consistently adds (or subtracts) an inch to the observed units. The final outcome is an overestimation (or underestimation) of the true population parameter. Unequivocally identifying whether a survey result is affected by response bias is not as straightforward as researchers would wish. Fortunately, research shows some conditions under which different forms of response bias can be found, and this information can be used to avoid introducing such biasing elements.

The concept of response bias is sometimes used incorrectly as a synonym of *nonresponse bias*. This use of the term may lead to misunderstanding. Nonresponse bias is related to the decision to participate in a study and the differences between those who decide to cooperate and those from whom data are not gathered. Response bias, on the other hand, takes place once a respondent has agreed to answer, and it may theoretically occur across the whole sample as well as to specific subgroups of the population.

Forms of Response Biases

Rather than being a direct function of respondents' own features per se, it is often the instrument (questionnaire item) characteristics that are responsible for this deviation, in the sense that there is something in the question or context that affects the way respondents undergo the cognitive process of responding, thereby distorting the true answer in some manner. This may occur consciously or not, and the resulting overreporting or underreporting may have different causes. A nonexhaustive list of different response biases is presented here to exemplify the kind of problems the researcher may encounter when conducting a survey.

1. Some effects are related to the length and type of the task. Aspects such as burdensomeness of the task may produce boredom or fatigue in the respondent, affecting the thoughtfulness of their answers.
2. The interaction between interviewer, respondent, and interviewing approach may also affect the way responses are produced. The most typical example of this is social desirability bias, but other effects might involve the interviewer's pace of speech, race, and gender. For example, fast-speaking interviewers may communicate to respondents that they are expected to give quick, off-the-top-of-their-heads answers. They may also affect how questions are understood.
3. The order of the questions and the order of response options may influence the likelihood of respondents to select a particular answer, eliciting context effects. The recency effect is an illustration of this type of bias; here, respondents are more likely to choose the last response options presented to them when the survey is conducted orally.
4. The wording of the question can tip the scale in one or another direction. Push polls are one example where the wording is intentionally manipulated with the intention to obtain a particular result, but unexpected or unintended wording effects are also possible.

5. Response styles are sometimes considered a type of response bias. When response options are presented, the way respondents use them may have a biasing effect. Some respondents seem to prefer a particular section of the scale as compared to others, irrespective of what their real attitude or behavior is.
6. Other forms of bias may appear as a result of lack of specificity in the definition of the task. Researchers need to be aware that conveying the need for high response accuracy is not a given. Discourse norms in everyday life dictate that precise estimates of certain quantities may be inadequate, unnecessary, or even undesirable. Spending several seconds trying to recall whether an activity lasted 13 or 14 days is usually not well received by the audience of a daily life conversation, where rough estimations are common. A similar phenomenon, the *rounding effect*, has been observed in surveys; researchers have identified that certain values (0, 25, 50, 75, and 100) are more likely to be chosen when using scales from 0 to 100.
7. People not only distort their answers because they want to create a positive impression on their audience, they may also edit (fake) their responses because they fear the consequences that the true answer might have and therefore do not wish to reveal the right information. Respondents, for instance, may lie about their legal status if they fear that confidentiality might be breached.

In essence, for any of the aforementioned examples, the reported information is not data the researcher is seeking. The distortion may set out at any stage of the cognitive processing: the comprehension of the question, the retrieval of the information, the judgment or the editing of the question. Question order effects, for example, can influence the way respondents understand the questions, while long reference periods can increase the likelihood of finding telescoping effects due to memory problems.

As with many other survey aspects, researchers can find that different cultures exhibit different patterns of response biases, which confound these with the actual substantive differences. Gender of interviewer bias, for example, may have a stronger presence in countries where the interaction of women with men outside their immediate family is rare or simply inappropriate.

Because of the widespread range of factors that can lead to response bias, all data collection methods are potentially at risk of being affected by response bias. Furthermore, different data collection modes may be more vulnerable to certain effects. Literature suggests, for instance, that self-administered modes reduce the impact of social desirability.

Similarly, different types of questions have also shown to be susceptible to response bias. Although attitude and behavioral questions receive perhaps more attention, demographic variables are also known to be sometimes biased. Household income, for example, is usually underreported among those with higher earnings and overreported by those with lower ones.

Avoidance Strategies

Researchers have proposed multiple ways to control and correct for response bias. Some of these strategies are broad ways to deal with the issue, whereas others depend on the specific type of effect. Conducting careful pretesting of the questions is a general way to detect possible biasing problems. An example of a specific strategy to detect acquiescence, for instance, is writing items that favor, as well as items that oppose, the object of the attitude the researcher wants to measure (e.g., abortion). Respondents that answer in an acquiescent way would tend to agree with statements in both directions.

If response bias is a function of how questions are worded, how they are ordered, and how they are presented to the respondent, then careful questionnaire design can help minimize this source of error. However, there are sources of error that cannot be predicted in advance and therefore are much more difficult to avoid. In that case, there are strategies that can be followed to identify response bias. If validation data are available, survey outcomes can be checked against them and response bias can be more precisely identified. Similarly, the use of split ballot experiments can provide insights about deviations across different conditions and point out biasing factors. Including interviewer characteristics and other external data to analyze survey outcomes can help reveal hidden effects. At any rate, the researcher should check for possible response biases in data, take bias into account when interpreting results, and try to identify the cause of the bias in order to improve future survey research.

Ana Villar

See also Acquiescence Response Bias; Context Effect; Interviewer-Related Error; Measurement Error; Primacy

Effect; Push Polls; Questionnaire Order Effects; Recency Effect; Record Check; Respondent Burden; Respondent Fatigue; Respondent-Related Error; Response Order Effects; Social Desirability; Split-Half; Telescoping; True Score

Further Readings

Belli, R. F., Traugott, M. W., Young, M., & McGonagle, K. A. (1999). Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring. *Public Opinion Quarterly*, *63*, 90–108.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Response Latency

Response latency is the speed or ease with which a response to a survey question is given after a respondent is presented with the question. It is used as an indicator of attitude accessibility, which is the strength of the link between an attitude object and a respondent's evaluation of that object. While response latency has been used for some time in cognitive psychology lab experiments, its use in surveys came about more recently. In telephone surveys, response latency is measured in milliseconds as the elapsed time from when an interviewer finishes reading a question until a respondent begins to answer.

There are four stages that survey respondents use when answering questions: (1) question comprehension, (2) retrieval of information from memory, (3) integration of the information to form a judgment, and (4) selection of an appropriate response option. Response latency measures the time it takes to retrieve, form, and report an answer to a survey question. Response latency data can provide much useful information about attitude accessibility that can be incorporated into data analysis. For example, when attitudes are modeled to predict subsequent behavior, respondents with more accessible attitudes (indicated by shorter response times) often exhibit a stronger relationship between the attitude and the subsequent behavior. Attitude accessibility as measured by response latency is just one way of measuring the strength of an attitude, but it can be consequential for attitude stability, a respondent's resistance to persuasion, as well as the influence of the attitude on behavior. Response latency has also been used as a method of pretesting survey questionnaires in order to identify problematic questions.

Response latency was first used in cognitive psychology lab experiments where the timer measuring the response latency is a function of the participants' own reaction time to a self-administered instrument. When adapted for use in telephone surveys, it is generally measured via a voice-activated or "automatic" timer (which requires special equipment) that senses when a response is given or through an interviewer-activated or "active" timer embedded into the programming of computer-assisted telephone interviewing (CATI) software. The active timer requires the interviewer to start and stop the timer at the appropriate time using the computer keyboard and then verify that the time measurement is valid. Response latencies are coded as invalid if the interviewer fails to apply the timer correctly or if the respondent asks for the question to be repeated. Response latency can also be measured using a "latent" or unobtrusive timer that is programmed into CATI software and is invisible to both interviewers and respondents. Latent timers simply measure the total duration of each question from the time the question appears on the interviewer's screen until the moment the respondent's answer is recorded. Such timers also can be used in computer-assisted personal interviewing (CAPI) and computerized self-administered questionnaires (CSAQ).

Regardless of how the response latency data are collected, the distribution of responses is frequently skewed, and the data require careful examination and cleaning before analysis. Invalid data and extreme outliers should be removed and the data transformed to eliminate the skew. Depending on how the data are collected, researchers using response latency data may also want to control for baseline differences among respondents in answering questions and interviewers in recording survey responses because some respondents are naturally faster in answering questions and some interviewers are naturally faster in recording responses.

J. Quin Monson

See also Attitudes; Attitude Strength; Comprehension; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Computerized Self-Administered Questionnaires (CSAQ); Outliers; Retrieval

Further Readings

Bassili, J. N. (1996). The how and why of response latency measurement in telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Answering questions* (pp. 319–346). San Francisco: Jossey Bass.

Huckfeldt, R., Levine, W., Morgan, W., & Sprague, J. (1999). Accessibility and the political utility of partisan and ideological orientations. *American Journal of Political Science*, *43*(3), 888–911.

Mulligan, K., Grant, J. T., Mockabee, S. T., & Monson, J. Q. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis*, *11*(3), 289–301.

Response Order Effects

A response order effect occurs when the distribution of responses to a closed-ended survey question is influenced by the order in which the response options are offered to respondents. Primacy and recency effects are two common types of response order effects. Primacy effects occur when response options are more likely to be chosen when presented at the beginning of a list of response options than when presented at the end. In contrast, recency effects occur when response options are more likely to be chosen when presented at the end of a list of response options than when presented at the beginning of the list. The research literature contains myriad examples of both types of effects.

Response order effects are typically measured by presenting different groups of respondents with a survey question with the response options in different orders and assessing the effects of order on the answer respondents give. For example, a random half of respondents in a survey might be asked, *Which do you think is more important for success in life: self-control or the ability to enjoy oneself?* The other random half of respondents would be asked, *Which do you think is more important for success in life: the ability to enjoy oneself or self-control?* A primacy effect would be observed if significantly more respondents answered "self-control" in response to the first question than in response to the second, but a recency effect would be observed if more respondents answered "self-control" in response to the second question than in response to the first. In questions with more than two categorical response options, the number of possible response option orders increases dramatically as the number of response options increases (e.g., there are 24 possible response option orders for a question with 4 response options). In questions with response options that fall along a scale (e.g., *How likely is it that you will watch the president's speech on television: extremely likely, very likely, somewhat likely, slightly likely, or not at all likely?*), the response options fall into a logical order. For these questions, response order effects can be assessed by providing half of respondents with the response options ordered in one direction (e.g., *extremely likely, very likely, somewhat likely, slightly likely, or not at all likely*) and providing the other half of respondents with the response options in the opposite direction (e.g., *not at all likely, slightly likely, somewhat likely, very likely, or extremely likely*).

A number of explanations have been provided for response order effects. For example, some researchers have argued that respondents have difficulty remembering all response options and that response order effects reflect the response options most memorable to respondents (those at the beginning and end of a list of response options). However, researchers have observed response order effects in very simple, short questions with only a few (e.g., two) response options. For these simple questions, it seems unlikely that respondents are unable to recall the question or the response options.

In a large body of evidence regarding response order effects in questions with categorical response options, recency effects have been observed in some cases, primacy effects have been observed in other cases, and in other cases, no significant response order effect was observed. Another theoretical account of response order effects provides an explanation for this mixture of findings. John Krosnick's satisficing theory suggests that although survey researchers hope respondents will answer questions by carefully and thoughtfully going through the four mental processes involved in answering survey questions (i.e., comprehending and interpreting the survey question, retrieving relevant information from memory, integrating that information into a judgment, and mapping their judgment onto the response format provided), respondents may not always be able or motivated to do so. Instead, they may shift their response strategies to minimize effort while providing a "satisfactory" response to the survey question (i.e., satisficing). In doing so, respondents are merely searching for strategies or cues in questions that they can use easily to find a satisfactory answer. One such strategy involves choosing the first response option that seems reasonable, and this strategy is believed to be responsible for response order effects.

Satisficing theory suggests that whether researchers observe primacy or recency effects in questions with categorical response options may depend on the mode in which response options are presented. When response options are presented visually, most respondents probably begin by considering the option presented first, then the second option, and so on. So if respondents choose the first reasonable response option they consider, primacy effects are likely to occur. But when response options are presented orally, respondents cannot think much about the first option they hear, because presentation of the second option interrupts this thinking. Similar interference occurs until after the last response option is heard, and at that point the last response option is likely to be the most salient and the focus of respondents' thoughts. People may also be most likely to remember the last response options in a long list of response options. So if respondents choose the first reasonable response option they consider, recency effects will occur. Consistent with this logic, mostly primacy effects have appeared in past studies that involved visual presentation of categorical response options, and mostly recency effects have occurred under oral presentation conditions.

In questions with response options that fall along a scale, however, mostly primacy effects have been observed, regardless of whether the response options are presented orally or visually. In questions with response options that fall along a scale, respondents who are reading or listening to the response options do not need to listen to the whole list of response options to form their answer to the question. Instead, they can infer the dimension on which they are being asked to make a judgment after just the first or second response option. For example, in the "likelihood" scale question described earlier, respondents are likely to know after just the first or second response option that they are being asked to report the likelihood of a particular behavior.

In addition to mode of presentation, satisficing theory posits that the strength of response order effects depend on three types of factors: (1) the respondent's ability, (2) the respondent's motivation, and (3) the cognitive difficulty of optimizing inherent in the question. Respondents with greater ability and motivation are less likely to satisfice. Satisficing is also more likely when (a) a question's stem or response choices are especially difficult to comprehend, (b) a question demands an especially difficult search of memory to retrieve needed information, (c) the integration of retrieved information into a summary judgment is especially difficult, or (d) translation of the summary judgment onto the response alternatives is especially difficult. Thus, recency effects in questions with orally presented, categorical response options are likely to be strongest among respondents low in ability and motivation and for questions that are more difficult.

Although evidence on the prevalence of response order effects suggests that researchers may want to estimate and control for such effects, there may be some cases in which this is not appropriate. In some cases, there may be norms about the order in which response options should be presented. For example, in questions with positive and negative response options (e.g., approve or disapprove), it is conventional to offer the positive response option first. Violating this convention may distract and confuse respondents and introduce error into their responses, thereby causing other types of potential measurement errors. So although in most cases, researchers may want to routinely rotate response order effects across respondents so that they can estimate and control for response order effects, this is most appropriate for questions without conventions about the order in which response options should be offered.

Allyson Holbrook

See also Closed-Ended Question; Primacy Effect; Recency Effect; Response Alternatives; Satisficing

Further Readings

Bishop, G., & Smith, A. (2001). Response-order effects and the early Gallup split-ballots. *Public Opinion Quarterly*, *65*, 479–505.

Holbrook, A. L., Krosnick, J. A., Carson, R. T., & Mitchell, R. C. (2000). Violating conversational conventions disrupts cognitive processing of attitude questions. *Journal of Experimental Social Psychology*, *36*, 465–494.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, *67*, 205–211.

McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research*, *20*, 60–103.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58–88.
Payne, S. L. (1949). Case study in question complexity. *Public Opinion Quarterly*, *13*, 653–658.
Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.

Response Propensity

Response propensity is the theoretical probability that a sampled person (or unit) will become a respondent in a specific survey. Sampled persons differ in their likelihood to become a respondent in a survey. These differences are a result of the fact that some persons are easier to get into contact with than are others, some persons are more willing and able to participate in a specific survey than others, or both. Response propensity is an important concept in survey research, as it shapes the amount and structure of unit nonresponse in a survey. The covariance between response propensity and the survey variable of interest determines the bias in survey estimates due to nonresponse. Theoretically, using response propensities, a researcher could entirely correct for nonresponse bias. However, such a correction requires that the researcher know the true value of this unobservable variable. In practice, researchers can use only estimates of response propensities, so-called propensity scores, using a logistic model that hopefully captures the concept well. The extent to which the nonresponse bias can be corrected using propensity scores depends on the quality of the propensity score model.

Determinants of Response Propensity

Response propensity is essentially a characteristic of the sampled person. Response propensities vary across persons according to their sociodemographic characteristics and various psychological predispositions. For example, persons in large households are easier to be contacted than those who live alone, whereas persons who are socially isolated are less likely to grant an interview than are people who are socially integrated. Response propensity can be regarded as a characteristic of the sampled person only, if defined as the probability that the sampled person will become a respondent in a random survey or, alternatively, as the probability

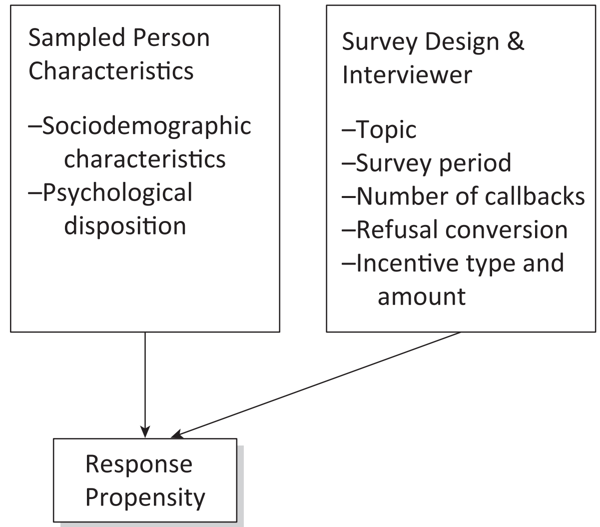

Figure 1 Determinants of response propensity

that the sampled person will become a respondent in a survey with "average characteristics." Considering a person's "baseline" response propensity within the characteristics of a certain survey then determines the probability that the person will be a respondent to that particular survey. So, in fact, response propensities are affected both by characteristics of the sampled person and by survey characteristics (see Figure 1).

Contactability and Cooperation

In the process of becoming a respondent, two important steps can be distinguished. The first step is that the respondent has to be contacted. The second step is that after being contacted, the respondent must be willing and able to cooperate with the surveyor's request. This observation suggests a straightforward decomposition of response propensity into two parts: contactability and cooperation.

The contactability $p_{contact,i,j}$ is the probability that individual i is contacted for survey j at any given moment in time. Typical characteristics of the sampled person that affect the contactability include factors such as (a) how often the person is at home, (b) whether there are physical impediments to contacting the person, and (c) to what extent the person is prepared to answer any incoming survey requests. Survey characteristics that affect contactability include (d) the number of contacts made by the interviewer, (e) the length of the survey period, and (f) the timing of the contacts.

For example, contact attempts in the evenings and on weekends are usually more successful than contact attempts at other times, thereby raising the response propensity.

The cooperation $p_{cooperation,i,j}$ is the probability that a contacted individual i will cooperate to survey j. Cooperation is in fact a conditional probability of cooperation on being contacted. From survey experience it is known that cooperation differs according to sociodemographic variables, such as gender, age, educational attainment, and the presence of children. The effects of sociodemographic variables are not believed to be causal. Theories on social isolation and authority have been developed to derive concepts that play a mediating role between these personal characteristics and cooperation. Many survey characteristics that affect cooperation, including the topic, the sponsor, incentives, perceived burden, and the interviewer, play a role in the decision of the contacted person to agree or refuse participation (see Table 1).

An interesting perspective on how the individual decision process of survey participation might work is provided by the leverage-saliency theory as articulated by Robert M. Groves and his colleagues. This theory acknowledges the threshold nature of survey participation. At the time of a survey request, a person will instantaneously evaluate different characteristics of the survey (e.g., topic, incentive, sponsor, burden) and will weight these characteristics on a personal scale. Depending on the "saliency" and "leverage" of the different characteristics of the survey, the scale will either tilt toward acceptance or toward refusal of the survey request.

Table 1 Survey characteristics that affect survey cooperation

Topic	Some survey topics are more popular than others.
Sponsor	Commercial organizations experience more refusals than official and academic research organizations.
Burden	Surveys with a high perceived burden (income and expenditure surveys, panel surveys) experience lower cooperation rates than surveys with low burden.
Incentives	Incentives offered to persons may stimulate cooperation.
Interviewers	Well-trained interviewers are more able to persuade persons to participate in the survey than are inexperienced interviewers. A persuasive counter to "I'm too busy" or "I'm not interested" is crucial to obtaining high cooperation rates.

Behind contactability and cooperation are two different processes that can have opposite effects. For example, in the case of a survey on cultural participation, culturally active persons may be harder to contact because of the fact that they are often away from home engaging in their cultural activities. At the same time, culturally active persons who are contacted may be very cooperative as their interest in the topic of the survey is above average.

For a given survey, contactability can be combined with cooperation to reveal the (overall) response propensity of sampled person i by multiplication of these probabilities:

$$p_{response,i} = p_{contact,i} \times p_{cooperation,i}$$

In survey research, next to contactability and cooperation, sometimes a third type of unit nonresponse is distinguished. It happens that persons who are contacted and who may even be willing to participate in the survey are not able to participate for reasons related to their physical health, literacy, or language problems. For surveys that deal with elderly and immigrants, this type of nonresponse can be serious and should not be ignored. Theoretically, it is possible to incorporate the inability to participate into a model for response propensity by extending the decomposition into three factors.

Using Response Propensity to Correct for Nonresponse Bias

In spite of surveyors' efforts to obtain high response rates (e.g., many callbacks, incentives, refusal conversion attempts), some sampled persons will become nonrespondents. It is common to apply post-survey adjustment techniques, such as post-stratification or raking methods, to reduce nonresponse bias. Alternatively, the concept of response propensity can be used to compensate for unit nonresponse, model it, and use estimated response propensities to correct for nonresponse bias.

The advantage of using the concept of response propensity over traditional methods is that this approach allows incorporating theoretical notions about survey participation into the procedure of bias reduction. From this perspective it makes sense to use a procedure that reflects the two-step process of survey participation and estimate multivariate logistic models for contactability and cooperation separately. The estimated models can then be used in a two-stage procedure to adjust the original sampling weight that corrects for unequal selection weights into weights to also correct for nonresponse bias. The quality of the procedure depends on the quality of the response propensity models. In practice, the quality of these models is still rather low. However, it is a challenge to find concepts that have strong relationships with contactability and cooperation, to obtain measures of these concepts on both respondents and nonrespondents, and to use these in estimating response propensity to correct for nonresponse bias.

Adriaan W. Hoogendoorn

See also Contactability; Contact Rate; Contacts; Cooperation; Cooperation Rate; Leverage-Saliency Theory; Nonresponse; Nonresponse Bias; Post-Stratification; Propensity Scores; Refusal Conversion; Response

Further Readings

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and illustration. *Public Opinion Quarterly*, *64*(3), 299–208.

Singer, E. (2006). Nonresponse bias in household surveys [Special issue]. *Public Opinion Quarterly*, *70*(5).

RESPONSE RATES

A response rate is a mathematical formula that is calculated by survey researchers and is used as a tool to understand the degree of success in obtaining completed interviews from a sample. In probability samples, where the intent of a survey is to project the results of the data onto a population (e.g., all adults in the United States), statistical theory rests on an assumption that data are collected from every unit, or person, selected. In practice, it is extremely rare for any survey to achieve this perfect level of cooperation from respondents. In turn, survey researchers may consider, examine, and when necessary, compensate for potential problems that this deficiency presents.

Response rates, sometime termed *outcome rates*, have traditionally been a topic of great interest because they describe the amount of nonresponse in a given survey. In doing so, they provide an indicator that can be used to better understand threats to the validity of survey data. Response rates inform researchers of the proportion of their sample that did not respond and also may lend insight into the reasons selected persons (or units) did not respond.

Background

Although nonresponse has been studied since the 1940s, serious efforts to standardize the measurement of nonresponse have arisen only within the last quarter of the 20th century. Furthermore, the common use of standardized response rate measurements has not yet been fully realized throughout the survey research profession.

Traditionally, there has been a great deal of overlap and inconsistency in both the definitions and formulas used to understand the concept of response rates. These discrepancies present a difficulty to the survey research profession because they often confuse consumers of survey information. Using consistent outcome rates is important because it allows the level of nonresponse to be compared more easily between different surveys. This provides researchers and clients or other end-users with a meaningful target when planning the design of research. Equally as important, standard outcome rates offer an important benchmark for understanding how well surveys performed.

For example, a lack of consistency prohibits the accurate comparison of nonresponse between two unique surveys, obscures agreement in target levels of nonresponse in research proposals, and hampers methodological research exploring nonresponse error.

In response to the historical differences among response rate calculations, the survey research profession has gradually worked toward a uniformly accepted set of formulas and definitions for nonresponse. These efforts are now spearheaded by the American Association for Public Opinion Research (AAPOR), which maintains a series of definitions, formulas, and dispositions that are continuously updated to reflect new technologies and changes in the survey research profession.

AAPOR Response Rates

AAPOR first published a series of response rates in 1998 for random-digit dialing and in-person surveys due to the concerted efforts of Tom. W. Smith and his colleagues. AAPOR based their development of the rates on the earlier work of the CASRO (Council of American Survey Research Organizations), which had published a set of formulas in 1982. Prior to that time, there had been numerous inquiries into the development of standards but no successful efforts put forth, at the association level, toward rate development.

Since the first release in 1998, AAPOR's volunteers have updated the response rates (and other outcome rates) three times (in 2000, 2004, and 2006). The most recent publication (2006) includes descriptions for calculating telephone, in-person, mail, and Internet survey response rates. AAPOR's development of response rates includes the formulas within a larger collection of "outcome rates."

Collectively, the four rates (response, cooperation, refusal, and contact) each help describe a different facet of survey nonresponse:

- *Response rates* describe the proportion of responders within a sample.
- *Cooperation rates* describe the proportion of responders who were contacted and who cooperated.
- *Refusal rates* describe the proportion of the sample who refused to take the survey.
- *Contact rates* describe the proportion of sample members who were contacted.

These rates also help describe the total nonresponse of a survey, as well as the type of nonresponse that a survey includes (e.g., refusal vs. noncontact).

Notably, AAPOR's set of outcome rates include numerous variations of each of the four types of formulas (response, cooperation, refusal, and contact). The six iterations that apply to response rates vary according to the type of information that is included in each part of the formula. For example, AAPOR Response Rate 1 (RR1) is calculated as follows:

AAPOR RR1:

$$\frac{\text{Completed Interviews}}{(\text{Completed Interviews} + \text{Partial Interviews}) + (\text{Refusals} + \text{Noncontacts} + \text{Other}) + (\text{Unknown Eligibility})}$$

AAPOR Response Rate 2 (RR2) is similar to RR1, except it considers partial interviews in the numerator of the formula. This potentially increases the response rate for a given survey.

AAPOR RR2:

$$\frac{(\text{Completed Interviews} + \text{Partial Interviews})}{(\text{Completed Interviews} + \text{Partial Interviews}) + (\text{Refusals} + \text{Noncontacts} + \text{Other}) + (\text{Unknown Eligibility})}$$

AAPOR's response rates are used widely within the survey research profession and academia. AAPOR maintains a policy of encouraging the use of their rates and definitions and has taken steps to see them proliferated throughout the survey research profession. Notably, at least two scholarly journals (*Public Opinion Quarterly* and *International Journal of Public Opinion Research*) have recognized the AAPOR formulas. Additionally, their use is endorsed by the CMOR (Council for Marketing and Opinion Research) for survey research conducted within the United States of America.

Survey response rates measure unit nonresponse. Unit nonresponse occurs when those who *have been selected* for participation in a survey *do not participate*. Nonresponders may not participate for numerous reasons ranging from circumstances where they plainly refuse to participate, to situations where they are never contacted and do not have the chance to participate. Survey nonresponse affects survey validity when nonresponders are different from those that do respond in ways that skew survey results. The following example illustrates a situation where nonresponders are absent from survey data in a way that affects survey validity:

> A researcher interviewed a sample of college students from ABC University about extracurricular activities. Those students who were more engaged in university activities also tended to participate in the university survey in greater proportions than other students. Those "active" students skewed survey statistics because of their very distinct feelings about extracurricular activities. This led the university to believe that the student body held a more favorable opinion of extracurricular activities than was actually the case.

Declining Response Rates

For many years, response rates have been declining in the United States (as well as in many other countries).

There are multiple factors that are believed to have contributed to this phenomenon. This trend of declining response rates is attributed to survey refusals and noncontacts.

Survey refusals are cases where a respondent receives a contact to complete a survey but declines to participate. The increase in survey refusals has been attributed to both social changes and reactions to changes within the survey research profession. These factors include (a) a growing expectation for privacy among the public; (b) the use of *pseudosurveys* as a guise to sell, fund-raise, push-poll, create marketing databases, or engage in political telemarketing; (c) the commoditization of research and substantial increase in the number of surveys being conducted; and (d) a decrease in the perceived value of surveys by society.

Noncontacts are situations where researchers are unable to communicate with the selected respondent. The growing problem of survey noncontacts has also had a wide breadth of contributing factors. Many of the situations that are thought to add to noncontacts vary largely across survey modes. For example, telephone research has been particularly susceptible to technologies affecting noncontacts such as the advent of telephone screening devices and services, cellular phone only households and number portability. However, other modes of recruitment or data collection are also challenged by unique circumstances that may magnify survey noncontacts (e.g., spam filters blocking Internet survey invitations, doormen preventing interviewer access to respondents in in-person surveys).

Additionally, response rates in the United States are subject to a third growing category of nonresponse: language. Many survey research organizations may want to interview persons who do not speak English and yet do not have the mechanisms in place to translate into other languages than English. If not addressed, this problem is likely to continue growing in scope.

Direction of Research on Nonresponse

Numerous studies have been conducted on response rate trends and the factors that may influence response rates for individual surveys. In more recent times, researchers have turned to studying the circumstances where survey nonresponse may be likely to pose threats to survey validity.

Nonresponse tends to be a complex, sophisticated phenomenon in survey research. As such, the meaning of response rates is often misinterpreted. It is important to view survey response rates in the context of the survey design.

Recent research in survey methods lends support to the idea that response rates must be considered along with other information. This convention contrasts somewhat with previous notions of researchers who believed a certain minimum response rate would offer sufficient protection (or mitigation) against nonresponse error, which is recognized nowadays to not be the case.

Evaluating Response Rates

The body of literature on response rates and survey nonresponse nevertheless indicates that response rates remain an important indicator of survey quality and should be considered when performing survey research. For this reason, it is recommended that response rates be calculated and considered when conducting survey research that relies on probability samples. However, it is also important to analyze response rates (and other outcome rates) in the context of the design of the study.

All available sources of information should be considered when exploring the meaning of response rates on a particular study. Important considerations for evaluating response rates may include the survey variables of interest, the survey population of interest and sample, survey design choices (e.g., use of incentives, timing, and nature of information given to respondents throughout survey process), and the design and administration of the survey instrument.

Patrick Glaser

See also American Association for Public Opinion Research (AAPOR); Contact Rate; Cooperation Rate; Council for Marketing and Opinion Research (CMOR); Council of American Survey Research Organizations (CASRO); *International Journal of Public Opinion Research* (IJPOR); Noncontact; Pseudo Polls; *Public Opinion Quarterly* (POQ); Refusal; Refusal Rate; Standard Definitions; Unit Nonresponse

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author. Retrieved April 10, 2008, from http://www.aapor.org/pdfs/standarddefs_4.pdf

Smith, T. W. (2002, August). *Establishing standards for survey research: The case of nonresponse*. Paper presented at the International Conference on Improving Surveys, Copenhagen.

Retrieval

Basic memory processes fundamentally affect the answers survey respondents give to survey questions, and retrieval is one of these memory processes. Retrieval refers to the active recovery of previously encoded information from long-term memory into conscious, working memory. Long-term memory that requires conscious retrieval is divided into (a) memory for facts and (b) memory for events. It is important for questionnaire designers to be aware of the limitations that retrieval processes place on the survey experience for the respondent.

The process of retrieving information and memories for events is closely related to the initial process of encoding, when a memory for the event is first stored by transferring information from active experience into long-term storage. The most difficult retrieval situation is unaided (free) recall, where there are no cues provided that relate to the desired information. However, questionnaires can avoid this situation by including cues in the survey questions that aid in recall. Retrieval is maximized when the cues present during encoding closely match the cues present in the questionnaire. If, for example, memory for a particular event is typically encoded in terms of time (i.e., filing your taxes), then survey questions that cue respondents to think about the event temporally will aid recall. The more cues there are, the more precise the retrieved information will be; for example, memory for a person's name may be aided by cueing specific events relating to that person, picturing their physical characteristics, thinking about the sound of her or his voice, or thinking about common acquaintances. Although the match between encoding and retrieval cues is important, it is also the case that, all other things being equal, some cues are generally better than others in aiding recall for events. Cues relating to the type of event work best; next best are cues relating to location and people and, finally, cues relating to time.

The depth of initial encoding is also important to later success in retrieving that information from memory; this phenomenon is referred to as the *levels-of-processing effect*. It is easier to remember information that was initially processed more deeply in comparison to information that was processed superficially. For example, information that is processed both visually (i.e., pictorially) and verbally is easier to remember than information processed in only one fashion.

Retrieval is greatly affected by the type of information that a survey question requires respondents to access. Memory access is first and foremost affected by how long ago the memory was initially encoded. Retrieval for older information and events is more difficult and error-prone, whereas memory for more recent information is easier to access. Certain kinds of information are also easier to retrieve from memory. For example, dates and names are forgotten easily. Information about a unique event is easier to remember, whereas memory for a specific event that is a repetitive and common experience of the respondent is more difficult to remember and more prone to errors. For example, it is easy to remember the details of meeting the president because this information is distinctive in comparison to other information in memory. In contrast, retrieving specific information about your last trip to the grocery store is more difficult to remember accurately because there are so many similar events in your past that the details for each event seem to blur into a general memory for "going to the grocery store." This difficulty leads to errors in identifying the source of a memory; that is, respondents are unable to accurately identify the specific originating event of the retrieved information.

It is important to note that the retrieval of information from long-term memory is a reconstructive process and fraught with possible error. When memory for an event or fact is accessed, this remembrance is not a perfect replica of that event or initially encoded fact. Retrieval errors affect survey responses in three common ways: (1) *Forgetting* occurs when respondents are simply unable to access information stored in memory, or the retrieved memory is partial, distorted, or simply false (false memories may occur even when respondents are quite confident in the accuracy and validity of the memory); (2) *estimation* occurs when respondents make a "best educated guess" that stands in for the retrieval of exact information, which often occurs because respondents satisfice; that is, they attempt to finish a survey by answering questions with as little effort as possible; and (3) *telescoping* occurs when respondents inaccurately include events that belong outside the reference period a survey question requires.

The constraints that retrieval processes impose on the ability of survey respondents to accurately recall information require special attention in the construction of survey questions. Questions should ask for information that is as recent as possible and provide cues that match the original encoding context. As noted earlier, cues relating to the type of event work best, followed by cues relating to location and people, and finally, cues relating to time. Cues that are distinctive and personally relevant to the respondent also aid retrieval. Lastly, questions that ask for information that is general or was initially superficially encoded are prone to greater error in comparison to questions with appropriate specific cues and therefore should be avoided if possible.

Gregory G. Holyk

See also Bias; Encoding; Questionnaire Design; Reference Period; Satisficing; Telescoping; Unaided Recall

Further Readings

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Reverse Directory

A reverse directory has two general definitions. The first is a residential telephone directory, which has been converted from a surname alphabetical ordered listing to a street name ordered listing. The second is a listing of addresses in a city, also known as a *city directory*.

A telephone directory covering a city or other geographic area consists of residential listed telephone numbers. The directory is in alphabetical order by surname. A first name or initials generally accompany the surname. The street address is then listed, although in some areas a household may decide to not have their address listed in the telephone directory. The telephone number follows the address. Because telephone directories are key-entered into a database by commercial firms, it is possible to manipulate the order of the listings.

In a reverse directory, the telephone directory database for a city is sorted into an alphabetical street order and then by address number within street name. In some cases the commercial firm will add additional information such as zip code, census tract, and census block number. This type of reverse directory makes it possible to sample households from very small geographic areas for a telephone survey or an in-person interview survey. However, there are several sampling issues that must be taken into account. One key issue is that the reverse directory is a listing of residential addresses with directory-listed landline telephone numbers, and therefore households with unlisted telephone numbers and households without landline telephone service are excluded. These types of households are likely to be demographically different from households that are listed in a reverse directory, thus leading to possible coverage error in a survey that is measuring the residential population in a local area.

The selection of a sample from a residential telephone directory, which has been converted from a surnamed alphabetically ordered listing to a street name ordered listing, involves the random selection of address listings. This can be accomplished by determining the number of pages in the directory and first selecting a systematic random sample of pages. For each selected page, one or more address listings can then be selected also through systematic random sampling. This approach assumes that the pages in the reverse directory contain approximately the same number of address listings. The survey researcher must keep in mind that such samples exclude telephone households with unlisted telephone numbers.

City directories, on the other hand, provide a listing of residential addresses in a city from which a sample can be drawn. RL Polk is the original commercial firm that sold city directories for many of the cities in the United States. The street guide section of the typical city directory provides a listing of streets in alphabetical order and addresses in numerical order. City directories are sometimes used to select dwelling units for area probability sampling of cities.

Sampling from city directories has its origins in area probability sampling, as discussed by Leslie Kish. Kish described sampling procedures for selecting an element sample of address listing from a city directory. For a large city, the selection of an element sample of address listings does not take advantage of two-stage sampling methods developed to reduce data collection costs for in-person surveys. This approach involves first selecting clusters of addresses and, for each sample cluster, drawing a sample of address listings.

Reverse directories are not used much anymore for sampling purposes. Today commercial sampling firms maintain computerized databases of residential directory-listed telephone numbers. Each listed telephone number is assigned to a county, census tract, block group and census block, as well as to a zip code. This makes it possible to sample residential directory–listed numbers from specific geographic areas, including small areas such as a single block group. For area probability sampling, the U.S. Postal Service (USPS) Delivery Sequence File (DSF) can be obtained for sample selection purposes. The DSF contains all delivery point addresses serviced by the USPS. Each delivery point is a separate record that conforms to all USPS addressing standards. Initial evaluations of the DSF as a means of reducing the costs associated with enumeration of urban households in area probability surveys have proven to be promising. Initial assessments of frame coverage found that the DSF covers approximately 97% of the households in the United States.

Michael P. Battaglia

See also Area Probability Sample; Directory Sampling; List Sampling; Probability Sample; Systematic Sampling; Reverse Directory Sampling; Unlisted Number; Unpublished Number

Further Readings

Iannacchione, V., Staab, J., & Redden, D. (2003). Evaluating the use of residential mailing addresses in a metropolitan household survey. *Public Opinion Quarterly*, *76*, 202–210.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Schejbal, J. A., & Lavrakas, P. J. (1995). Coverage error and cost issues in small area telephone surveys. *American Statistical Association 1994 Proceedings: Section on Survey Research Method* (pp. 1287–1292). Alexandria, VA: American Statistical Association.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Reverse Directory Sampling

A reverse directory is a residential telephone directory that has been converted from a surnamed alphabetically ordered listing to a street name ordered listing. Reverse directory sampling refers to the selection of a sample of address listings—which may be residential, nonresidential, or both—from a reverse directory. The selection of a sample from a reverse directory involves the random selection of these address listings. This can be accomplished by determining the number of pages in the directory and first selecting a systematic random sample of pages. For each selected page, one or more address listings can then be selected also using systematic random sampling. This approach assumes that the pages in the reverse directory contain approximately the same number of address listings.

Today, such manual sampling procedures are rarely used. Telephone directories are key-entered so that computerized databases of telephone directories are available through various commercial sources. Geographic codes (such as county FIPS [Federal Information Processing Standard] code), census tract number, block group number, and zip code are assigned to each telephone number in the database. This makes it possible to draw a random sample of telephone numbers from a specific geographic area (e.g., a single zip code). The survey researcher must recognize that such samples exclude telephone households with unlisted telephone numbers and households without a landline and therefore may contain considerable coverage error.

Michael P. Battaglia

See also Reverse Directory; Unlisted Number; Unpublished Number

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Reverse Record Check

In survey research, a reverse record check is a method that can be used to verify the accuracy of data that a respondent provides against an information source that contains the “true” answer. Use of this method is not always feasible, even with ample funding, because external sources of information against which a survey response can be validated often do not exist.

The effect of human memory on the response to the survey questions has been the concern of survey methodology for many years. A survey respondent has to be in the position of providing accurate answers to survey questions involving memory. Naturally, this must depend on the nature of questions, but a number of empirical studies have demonstrated that fallibility of memory can be an important source of measurement error. In addition to memory failures, there are other reasons that the answers respondents sometimes provide are inaccurate, thus contributing to measurement error.

Perhaps the most common method of assessing measurement error is the use of a record check study or validation study. Such a study generally assumes that information contained in the external records is without error, that is, that the records contain the true values on the survey variables. In fact this may not be true, as the records may themselves be inaccurate, incomplete, or both. Furthermore, the matching and finding of records for survey individuals is often troublesome and expensive. Errors may occur in matching an individual's survey answer(s) and record data, and if no attempt is made to reconcile differences, such mismatches may indicate response errors where none exists. There are three kinds of record check study designs: (1) reverse record check study, (2) forward record check study, and (3) full design record check study. The reverse record check study is generally based on a retrospective design in which the entire sample is drawn from the record file for persons with a trait under study, interviews are conducted containing questions about information also contained on the records, and a comparison of survey data with record data is performed to estimate the extent and nature of any measurement error. The term *reverse record check* is used because after the survey is completed, the researcher goes back to the records to check the survey responses. The information can be gathered by using so-called retrospective questions, referring to a certain period of time preceding the date of the interview. In retrospective surveys, many types of recall errors, such as omission, telescoping (forward or backward), confusion, and reporting load effects, can occur.

Earlier studies involving reference period questions showed forward bias in reporting. When memory is inexact, forward bias will arise in answering such questions. Even if an upper bound is not imposed by the investigator, subjects may impose such a bound on their own reports. If subjects bound their reports, it leads to truncation of the distribution of reports just as when bounds are imposed in the question; it also leads to forward bias in reporting. It is important to note that reverse record check studies by themselves fail to measure errors of overreporting. Instead, reverse record checks can measure what portion of the records sampled correspond to events reported in the survey.

Record linkage and alternative matching procedures can be compared with the perfect matching situation. The expected *event location matching* is achieved by taking the record time as the base for the matching operation. Through this approach, reported events are matched with their expected location in the corresponding records. By this procedure, telescoping effects are replaced by omissions for given locations. One of the major drawbacks of this approach is the difficulty in placing the expected event location realistically. A second approach is nearest distance event matching, where the reported events are matched with their nearest counterpart in the recorded events in the ordered set for each respondent's individual records. This approach creates larger deviations between the reported and recorded values for each individual's record, because omission is replaced by telescoping effects in some locations. By this approach, omissions will not be eliminated and will still appear in the data set at a later location, when all the events are matched. Matching procedures conducted under different assumptions also can create different patterns in the matched data for the reverse record check study.

H. Öztas Ayhan

See also Errors of Omission; Event History Calendar; Measurement Error; Nonsampling Error; Overreporting; Record Check; Telescoping

Further Readings

Ayhan, H. Ö., & Isiksal, S. (2004). Memory recall errors in retrospective surveys: Reverse record check study. *Quality & Quantity 38*(5), 475–493.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Sudman, S., & Bradburn, N. M. (1973). Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association*, *68*, 805–815.

ρ (Rho)

A statistic that quantifies the extent to which population units within clusters are similar to one another (i.e., the degree of homogeneity within clusters) is called the *intraclass correlation coefficient* (ICC) and is often denoted by the Greek letter *rho* (ρ).

When population units are grouped or clustered into larger units, which are themselves easier to identify and sample (i.e., children grouped into classrooms or elderly citizens grouped into nursing homes), one- or two-stage cluster sampling becomes an appealing, cost-effective, and practical choice for a sampling strategy. These benefits are often counterbalanced by the usual expected loss in efficiency and precision in estimates derived from cluster samples that is, in large part, due to the fact that units within clusters tend to be more similar to each other compared to units in the general population for many outcomes of interest.

The computation of ρ essentially provides a rate of homogeneity for elements within a cluster relative to the overall population variance, as seen by the following equation:

$$ICC = \rho = \frac{\sum_{i=1}^{C}\sum_{j=1}^{M}\sum_{k \neq j}^{M}(y_{ij} - \bar{y}_{iU})(y_{ik} - \bar{y}_{iU})}{(CM-1)(M-1)S^2}, \qquad (1)$$

where C is the number of clusters in the population, M is the number of elements within each cluster, y_{ij} is the measurement for the j^{th} element in the i^{th} cluster, y_{ik} is the measurement for the j^{th} element in cluster $i, \bar{y}_{iU}$ is the population mean for the i^{th} cluster, and S^2 is the finite population variance defined by

$$S^2 = \sum_{i=1}^{C}\sum_{j=1}^{M}\frac{(y_{ij} - \bar{y}_U)^2}{CM-1}, \qquad (2)$$

where $\bar{y}_U$ is the population mean.

Note that Equation 1 is equivalent to a simpler formula containing values easily obtained from an ANOVA table, accounting for clustering as follows:

$$ICC = \rho = 1 - \frac{M}{M-1}\frac{SSW}{SST}, \qquad (3)$$

where $SSW = \sum_{i=1}^{C}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{iU})^2$ is the sum of squares within clusters and $SST = \sum_{i=1}^{C}\sum_{j=1}^{M}(y_{ij} - \bar{y}_U)^2$ is the total sum of squares about $\bar{y}_U$.

From Equation 3 and the fact that $0 \leq SSW/SST \leq 1$, it follows that $\frac{-1}{M-1} \leq \rho \leq 1$. If there is complete duplication within each cluster, then the ICC takes on the highest possible value of 1 to indicate complete homogeneity within clusters; on the other hand, if the heterogeneity within clusters is consistent with that of the overall population, then the ICC will assume its smallest value of $-1/(M-1)$. Cluster sampling will be more efficient than simple random sampling with the same overall sample size whenever $-1/M-1 \leq \rho \leq 0$ and less efficient when the ICC values are positive and closer to 1.

Consider the following example to illustrate the computation of the ICC. Researchers are interested in determining the average fruit and vegetable intake of staff members of a nationally franchised health club in preparation for a campaign to promote exercise and diet among its members. The population consists of five franchised health clubs that each have eight staff members. The fruit and vegetable intake for each population member is provided in Table 1. In this scenario the number of clusters is five ($C = 5$), and the number of elements per cluster is eight ($M = 8$).

From the SSW and SST obtained with these results the ICC is computed using Equation 3 as follows:

$$ICC = \rho = 1 - \frac{8}{8-1} \times \frac{5.932}{114.690} = 0.941.$$

This large, positive ICC indicates that, on average, the staff members within each cluster tend to consume similar amounts of fruits and vegetables per day. In other words, the clusters are extremely homogeneous. The homogeneity within clusters also means that a one-stage cluster sample would be less efficient than a simple random sample of the same size (i.e., the design effect [*deff*] would be greater than 1).

It should be noted that in the example, population values were provided. In practice, estimated ICCs are obtained using comparable sample statistics, such as sample variance and within and total sums of squares derived from sample data.

Other applications of ρ include its use in the following: (a) determining the necessity of analyzing data using a multi-level modeling technique; (b) evaluating longitudinal data; (c) assessing intercoder reliability in research with complex coding schemes where the Kappa statistic would not suffice; and (d) studying interviewer-related measurement error due to the idiosyncratic effects that interviewers likely have on the subset of individual interviews each of them completes.

Trent D. Buskirk and Sarah Shelton

Table 1 Population values of fruit and vegetable intake for five health club franchises

Staff Member	*Cluster*	*Average Daily Fruit and Vegetables*	*Cluster Average*	$(y_{ij} - y_{iU})^2$	$(y_{ij} - y_U)^2$
1	1	2.05	2.35	0.090	2.250
2	1	2.20		0.023	1.823
3	1	2.15		0.040	1.960
4	1	2.30		0.003	1.563
5	1	2.45		0.010	1.210
6	1	2.50		0.023	1.103
7	1	3.05		0.490	0.250
8	1	2.10		0.063	2.103
1	2	4.00	4.03	0.001	0.203
2	2	3.75		0.078	0.040
3	2	4.10		0.005	0.303
4	2	4.35		0.102	0.640
5	2	4.05		0.000	0.250
6	2	3.90		0.017	0.123
7	2	4.00		0.001	0.203
8	2	4.05		0.000	0.250
1	3	3.35	3.38	0.001	0.040
2	3	3.50		0.014	0.002
3	3	3.65		0.073	0.010
4	3	3.20		0.032	0.123
5	3	3.10		0.078	0.203
6	3	3.40		0.000	0.023
7	3	3.45		0.005	0.010
8	3	3.40		0.000	0.023
1	4	6.10	6.40	0.090	6.503
2	4	7.10		0.490	12.603
3	4	6.25		0.023	7.290
4	4	5.95		0.203	5.760
5	4	7.05		0.422	12.250
6	4	6.40		0.000	8.123
7	4	6.05		0.123	6.250

Staff Member	*Cluster*	*Average Daily Fruit and Vegetables*	*Cluster Average*	$(y_{ij} - y_{iU})^2$	$(y_{ij} - y_U)^2$
8	4	6.30		0.010	7.563
1	5	0.50	1.61	1.232	9.303
2	5	0.75		0.740	7.840
3	5	1.25		0.130	5.290
4	5	1.65		0.002	3.610
5	5	2.30		0.476	1.563
6	5	2.25		0.410	1.690
7	5	2.05		0.194	2.250
8	5	2.10		0.240	2.103
		$\bar{y}_U = 3.553$	Total:	5.932	114.690
				SSW	SST

See also Analysis of Variance (ANOVA); Cluster Sample; Design Effects (*deff*); Interviewer-Related Error; Simple Random Sample

Further Readings

Fowler, F. J., Jr. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, *57*, 92–115.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrences*. New York: Oxford University Press.

ROLE PLAYING

Role playing is an educational technique that is used in the training of survey interviewers. It involves face-to-face or telephone interviewers practicing the tailored use of the introduction of the survey in which an attempt is made to gain cooperation from a respondent, and practicing the proper way to administer the questions to respondents, or both. Role playing generally takes place toward the end of training, after the interviewers have been methodically exposed to the entire questionnaire by the trainer, on a question-by-question basis. The role playing part of training may last as long as 2 hours depending on the length and complexity of the questionnaire. Interviewers generally enjoy the role playing part of training, in part because it allows for their active participation in the training process.

Role playing typically takes two forms. One is where the supervisor or trainer takes on the persona of a respondent and interviewers in the training session go through the introduction or the questionnaire, or both, in a round-robin fashion asking the questions to the trainer and going on to the next appropriate question depending on the answer just given by the trainer. The trainer often varies his or her persona while this is happening—sometimes being cooperative and other times being uncooperative when the introduction is being practiced, and sometimes being an "easy" respondent and others times being a "difficult" respondent when questions are being asked. The second form of role playing is where interviewers are paired up and take turns interviewing each other with the questionnaire. As this is taking place, supervisory personnel typically move around the room observing the practice, making tactful suggestions for improvement where appropriate and being available to answer questions that might arise.

Role playing also is used in this "mock interviewing" fashion to help newly hired interviewers learn various basic interviewing skills. It also is used to help experienced or senior interviewers learn advanced interviewing techniques, such as with refusal avoidance training.

Paul J. Lavrakas

See also Interviewer-Related Error; Interviewer Training; Refusal Avoidance Training (RAT); Tailoring

Further Readings

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing*. Newbury Park, CA: Sage.

Stewart, C. J., & Cash, W. B. (2007). *Interviewing: Principles and practices* (12th ed.). New York: McGraw-Hill.

Tarnai, J., & Moore, D. L. (2008). Measuring and improving telephone interviewer performance and productivity. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 359–384). New York: Wiley.

Rolling Averages

Rolling averages, also known as moving averages, are a type of chart analysis technique used to examine survey data collected over extended periods of time, for example, in political tracking polls. They are typically utilized to smooth out data series. The ultimate purpose of rolling averages is to identify long-term trends. They are calculated by averaging a group of observations of a variable of interest over a specific period of time. Such averaged number becomes representative of that period in a trend line. It is said that these period-based averages "roll," or "move," because when a new observation is gathered over time, the oldest observation of the pool being averaged is dropped out and the most recent observation is included into the average. The collection of rolling averages is plotted to represent a trend.

An example of how rolling averages are calculated is as follows. Imagine that a survey analyst is interested in computing 7-day rolling averages (i.e., a 1-week period) for a period of 52 weeks (364 days, approximately 1 year). Let us assume that the variable of interest was measured daily over a period of 52 weeks. The analyst would have to consider the simple average of Day 1 to Day 7 as the first rolling average, represented as $RA_1 = (d_1 + \cdots + d_7)/7$. The second 7-day rolling average would be the average of Day 2 to Day 8, represented as $RA_2 = (d_2 + \cdots + d_8)/7$. Subsequent rolling averages would be $RA_3 = (d_3 + \cdots + d_9)/7, \cdots$, $RA_{358} = (d_{358} + \cdots + d_{364})/7$. The analyst then would have 358 points or rolling averages to plot a trend across the 364-day year.

In general, simple rolling averages are calculated as $RA_t = \sum_{i=1}^{k} d_i/(k-i)+1$, where RA_t represents the set of rolling averages for specific time periods (t), d_i represents one unit in the rolling average, and $(k-i)+1$ is the total number of time points in the rolling average (e.g., in the 7-day rolling averages example, $k=7$). Overall, it is up to the analyst to decide the total number of time points, $(k-i)+1$, to be averaged. For example, when variations are expected to occur within a 1-week period, the analyst can select 3-, 5-, or 7-day rolling averages, whereas in studies whose variations occurs monthly, 15-, 28-, or 30-day rolling averages could be selected. In studies with a larger scope, 1-, 2-, 3-year, or longer rolling averages would be needed.

Rolling averages reduce short-term effects; as a consequence, variations across time are decreased and the direction of the trend is more readily clarified. In that sense, variables subject to seasonality or periodic fluctuations in long-term studies are conveniently represented as rolling averages. For example, results of daily pre-election surveys conducted under the same methodology and by the same polling agency tend to fluctuate frequently because of campaigns, media-related effects, or any other aspect around elections; thus, a way to reduce variability and emphasize a voting-intention trend would be by means of rolling averages. Other variables subject to seasonality are, for instance, measures of donation behavior in fund-raising studies or measures of exercise levels in health fitness surveys. These variables would be likely to display noticeable peaks or dips in winter and summer months.

Rolling averages do not have inherent predicting properties—they are mainly used to provide a more accurate idea of the construct being monitored over time by reducing variations due to temporary events.

Nevertheless, in situations where the long-term trend is relatively stable, and after careful examination of other indicators, rolling averages may allow the analyst to foresee upcoming observations. In addition to the already explained simple rolling averages, there are various types of rolling averages such as weighted, exponential, triangular, and variable rolling averages. While the simple version described in this entry assigns equal weight to all observations, weighted and exponential rolling averages tend to give greater weight to recent observations. Triangular rolling averages assign greater weight to observations in the middle of the group, and variable rolling averages assign weight depending on the level of variability to be reduced.

René Bautista

See also Tracking Polls

Further Readings

Nebraska Department of Environmental Quality. (2002). *How to calculate yearly and rolling totals and rolling averages*. Retrieved May 18, 2008, from http://www.deq.state.ne.us/Publica.nsf/c4afc76e4e077e11862568770059b73f/4cef88e8eab6bcf105256c55007535b7?OpenDocument

Roper, Elmo (1900–1971)

Elmo Roper, born in 1900, was a leader in the fields of public opinion polling and market research for more than 30 years. He did not have formal academic training in these fields and, in fact, did not finish college. Rather, as the co-owner of a jewelry store in the 1920s, he became interested in customer opinions. He eventually left the jewelry business, and in 1933 he became a co-founder of one of the country's first market research firms, Cherington, Roper and Wood. Five years later Roper left that firm to found Roper Research Associates, Inc. He retired from the firm in 1966 but remained a senior consultant until his death in 1971.

Roper was deeply involved in popularizing market research and public opinion polling and increasing the influence of the fields in both private industry and within the U.S. government. From 1935 to 1950 he was the director of the "Fortune Survey" of public opinion, conducted for *Fortune* magazine. This was the first national opinion poll conducted using scientific sampling strategies.

In 1936 Roper solidified his reputation when the Fortune Survey very closely predicted the results of the presidential contest between Franklin Roosevelt and Alf Langdon. Roper bested the other major polls at the time, the Gallup Poll and the Crossley Poll. Gallup and Crossley predicted Roosevelt would win with 50% to 55% of the vote, while Roper predicted a Roosevelt win with more than 60% of the vote. Roosevelt won the election with 60.7% of the vote. This accuracy helped to establish scientific polling's position and importance on the political landscape. Roper maintained a high profile for the next 30 years. He accurately predicted the next two presidential elections, in 1940 and 1944, predicting results that were within 1% of the actual vote.

In 1948 he, like nearly all other pollsters at the time, incorrectly predicted that Thomas Dewey would defeat Harry Truman in the race for president. The field was badly shaken, but Roper's business survived, with some changes. Election polling in 1948 had stopped weeks before the election because pollsters believed the nation's voters would not change their minds close to the election. Election polling is now conducted up to the election itself and, in the case of exit polls, during it. Probability sampling also replaced quota sampling as the dominant methodology after 1948.

Roper Research Associates conducted polls for various government agencies and many private companies. Roper conducted polls for the New York Stock Exchange and the CBS television network. He was a deputy director of the Office of Strategic Services during World War II. He also wrote a syndicated newspaper column, had a radio show, and was editor at large for the *Saturday Review* magazine.

Outside the world of polling, Roper was a liberal democrat involved in many causes of the day. He sat on the board of Planned Parenthood. In 1955 he was elected president of the Atlantic Union Committee, succeeding Associate Justice Owen J. Roberts of the United States Supreme Court. The group sought to strengthen ties between Canada, England, Belgium, the Netherlands, and France as a counterbalance to the Soviet Union.

Roper believed that public opinion polls were important tools in a democracy and felt that good public opinion research could help scholars and government officials make more informed decisions. To this end, he believed that public opinion data from a variety of sources should be stored in a permanent collection, and in 1946 he created the Roper Center for

Public Opinion Research at Williams College. The center included data not only from Roper's work but also from the Gallup Poll and the Crossley Poll, among others. The Roper Center moved to the University of Connecticut in 1977 and remains the world's foremost archive of public opinion data.

Eric White

See also Probability Sample; Quota Sampling; Roper Center for Public Opinion Research

Roper Center for Public Opinion Research

The Roper Center for Public Opinion Research, which now is located at the University of Connecticut, and is one of the world's largest collections of data on the social sciences, was founded at Williams College in 1946 by Elmo Roper, who anticipated the future value of the studies he and other survey research pioneers were conducting to better understand what was on the collective minds of the public. His vision was of the academic scholar in pursuit of the thoughts, opinions, and behaviors of the common person during the early days of polling. Shortly thereafter he was joined by George Gallup, Sr., who contributed his 17 years of data, followed by Archibald Crossley, the National Opinion Research Center, Opinion Research Corporation, and the earliest state polls from Minnesota, Texas, and Iowa. More than six decades later, the Roper Center archives are the largest archives of survey data in existence, including the work of more than 250 survey organizations, representing thousands of studies and hundreds of thousands of questions asked in the United States and abroad. A list of the center's data contributors reads like a "who's who" of the profession, including the major news media—paper and electronic—that sponsor and conduct surveys on a multitude of topics.

This comprehensive collection has fulfilled Elmo Roper's dream that the data be preserved for future generations. However, today the center's role of data steward goes beyond preservation measures and has expanded to include the provision of easy access tools to facilitate use of the collection. The center's membership includes nearly every major university in the United States and several in Europe and elsewhere.

The Tools

The iPOLL database contains a half-million questions and responses from surveys conducted in the United States since 1935. Using simple search terms, it is possible to identify hundreds of relevant questions on most public policy and social issues in a matter of seconds. Of these half-million questions, about 60% are from surveys where the entire data set has been archived with the Roper Center; the other 40% are from surveys that have yet to arrive, are only available in report form, or are archived elsewhere. Using the integrated RoperExpress service, it is possible to move directly from a question in the iPOLL database to the documentation and download the data set, as well. This service permits researchers to conduct secondary analyses (e.g., cross-tabulations that reveal more details about respondents answering in a particular way) on their desktop computers.

The drill-down functionality that permits this on-demand download of the documentation and data files is distinctive to the Roper Center's data archives. Although systems exist that permit searching at the question level, most of those organizations do not maintain the raw data set files. Conversely, there are institutions that offer direct download of data set files but have not developed a question-level retrieval system.

Contributions to the Field

To improve the understanding of the public in the aggregate, the Roper Center partners with those who conduct survey research. To achieve this, it has developed systems that permit broad access and utilization of the data entrusted to its care. Additionally, the Roper Center Web site offers educational modules to acquaint researchers with the fundamentals of opinion polling and to assist with interpreting results.

There are two important ways in which the archives serve the survey research profession. The first is educational. Sound secondary data analysis relies upon principles of triangulation—examining multiple sources of data and contextual variations in question wording, timing, or methodology for the purpose of building a firm base for understanding a complex phenomenon. This is more than gathering numerous data points;

it brings together different "views" into the topic by using an array of sources. The comprehensive iPOLL database decries the notion of cherry-picking data by providing broad, easy access to the data. Survey practitioners are further served by these resources when they observe the work generated by their contemporaries. Pointless "reinvention of the wheel" is avoided, and worthwhile research can be repeated.

The center also promotes the intelligent use of public opinion data by providing educational modules on survey practices and introductory views into different topics. Its Web site offers a glossary, online analysis tools, and training tutorials. Two teaching modules, dubbed Polling 101 and Polling 201, cover sources of error, sampling, reading data tables, and select criteria for assessing survey quality and strategies for analyzing data. The content is straightforward and presented clearly, inviting to both the novice and the professional seeking a refresher on survey research methods and terminology.

The second way in which the archives contribute to the profession is by providing a safe venue to ensure its data will exist in usable formats in perpetuity. Unlike Web sites that make data sets available only temporarily, the Roper Center carefully processes each data set and integrates the data into its permanent archives, migrating to current formats and creating finding-aids like iPOLL and RoperExpress to make such research less complicated.

Furthermore, this venue permits researchers from various sectors to scrutinize the surveys. Critics and supporters alike are provided the opportunity to test theories by easily accessing the individual-level data files to run secondary analyses and statistical tests, which makes the industry stronger as a whole. Disclosure has always been important in this field, and quality survey organizations have believed that depositing data with an institutional archive that permits such investigation can reveal ways to improve upon their work.

Lois E. Timms-Ferrara

See also Roper, Elmo

Further Readings

Rockwell, R. C. (2005). The role of the Roper Center in the polling industry. In S. J. Best & B. Radcliff (Eds.), *Polling America: An encyclopedia of public opinion*. Westport, CT: Greenwood Press.

Rockwell, R. C. (2001). Data archives, international. *International encyclopedia of the social and behavioral sciences*. Paris: Elsevier.

Roper Center for Public Opinion Research: http://www.ropercenter.uconn.edu

Smith, T., & Forstrom, M. (2001). In praise of data archives: Finding and recovering the 1963 Kennedy Assassination Study. *IASSIST Quarterly*, *25*(4), 12–14.

Smith, T. W., & Weil, F. D. (1990). A report: Finding public opinion data: A guide to sources. *Public Opinion Quarterly*, *54*(4), 609–626.

ROTATING PANEL DESIGN

A rotating panel design is a survey sampling strategy sometimes used when estimates are produced regularly over time. Under such a design, equally sized sets of sample units are brought in and out of the sample in some specified pattern. These sets, often called *rotation groups*, may be composed of households, business firms, or other units of interest to the survey. Rotating panel designs are used to reduce the variances of estimators of level or change and often to reduce the survey costs associated with introducing a new unit into the sample. The wide variety of such designs reflects (a) the types of data and the estimates to be produced by the survey, (b) the statistical relationships among the characteristics of interest, (c) the operational costs of the survey, (d) the burden on respondents, and (e) the effects of multiple contacts on data quality.

Examples of Rotating Panel Designs

Under one type of rotation scheme, during each time period, one rotation group is canvassed for the first time, while another is canvassed for the final time. As an example, in the Labour Force Survey conducted by Statistics Canada, the sample of households is divided into six rotation groups, with a new group entering each month. Thus, in any month, one group is canvassed for the first time, one for the second time, and so forth. After 6 months of responding, households are retired from the sample. The Labour Force Survey conducted by the Australian Bureau of Statistics uses a similar scheme, dividing the sample into eight rotation groups; households are in sample for 8 months.

The rotation pattern can be more complex. The Current Population Survey, jointly sponsored by the U.S. Bureau of Labor Statistics and the U.S. Census

Bureau to measure the U.S. labor force, uses eight rotation groups. A household is interviewed for 4 consecutive months, temporarily dropped from the sample for 8 months, brought back into sample for 4 more months, then retired. In any month one rotation group is interviewed for the first time, one for the second time, and so on, while one group is interviewed for the eighth and final time. For each of the surveys mentioned, the rotation pattern is said to be a "one-level" design because the respondents report for only the current period of time (1 month).

Other examples of surveys that use a rotating panel design are the National Crime Victimization Survey and the Consumer Expenditure Survey, each conducted by the U.S. federal government. In the Census Bureau's Survey of Income and Program Participation, although cross-sectional estimates are produced, the people in sample are contacted in a rotating pattern to obtain longitudinal estimates, that is, estimates of gross changes over time. These three surveys are examples of multi-level designs, as respondents report for several prior periods (months) during the same interview.

Under a different type of rotation scheme, groups cycle in and out of sample indefinitely or at least until the entire sample is retired. A good example is the former (before 1998) design of the monthly surveys of retail and wholesale trade in the United States, where the noncertainty sample was divided into three rotation groups. In each of 3 consecutive months, only one of the three groups was canvassed. This pattern was repeated, with each group reporting every third month, for the 5-year life of the sample.

Considerations in Selecting and Implementing the Design

In many ways, a rotating panel design is a compromise between taking a new, independent sample for each period and canvassing the same sample units repeatedly, that is, a complete sample overlap. Each of these extremes has its advantages statistically and in practice; a rotating panel design, sometimes thought of as a partial sample overlap, tries to reap the benefits of each.

Statistically, if the correlations in variables of interest are strong over time, estimates of change can usually be made more accurate by overlapping some or the entire sample across time. This is the case even without applying composite estimation, a statistical method that combines data from several sources, for example, different periods of time. However, composite estimation can reduce variances further in many cases, especially where estimates of means or totals are desired.

Many factors are considered before implementing a rotating panel design. Of primary concern are the types of data and estimates to be produced and how they rank in importance. The relative importance of estimates of level (totals, means, and proportions) at specific points in time, change across periods, and averages over time must be weighed. The variances and biases of these estimators will typically depend on the type of design, the amount of rotation, and the correlations among the rotation group estimates over time.

Operational aspects, such as survey costs, the mode of data collection, and respondent burden, are important considerations as well. For surveys that are conducted repeatedly, much of the cost may be incurred the first time a sample unit is contacted. For example, canvassing a region to devise an area sample can be very expensive, especially if the units are interviewed only once. Even with list sampling, the tasks may include (a) preparing introductory questionnaire materials, (b) finding the household or establishment (for a personal visit), (c) explaining the survey or special procedures to the respondent, (d) filing background information about the sample unit, and (e) preparing the database for each unit. Many of these steps need not be repeated if the same unit is canvassed a second or third time, as in a rotation design. For example, in some household surveys, the first contact is made in person because a telephone number is not available. When a rotating panel design is used, subsequent contacts can often be completed by telephone, reducing the cost per interview.

Issues of respondent burden and response are relevant in a rotating panel design. The burden on an individual household or establishment typically increases with the number of interviews and the length of the questionnaire. With increased burden might come decreased response rate or lower-quality data. For household surveys, how to handle people who move and what effect they have on the estimates can be a problem. These issues become more prominent as the time between contacts with the sample unit increases. For business surveys, various approaches have been proposed for continuous sample selection while controlling the overlap over time and the burden. The *permanent random numbers technique* has been implemented in several countries, whereby a unique random number is assigned to each sampling unit and stays assigned to that unit as long as the unit is in the sampling frame. This number is used to reduce the chances that any given unit will be selected for interviewing in two consecutive waves of the survey, thereby minimizing respondent burden over the entire panel study.

Other problems can arise with a rotating panel design. As the respondents in the sample grow accustomed to the repeated interviews, the values of their responses may be affected, a consequence known as panel conditioning. Specifically, time-in-sample bias refers to the deviation in the response for a rotation group that is interviewed for the first (or second, or third, etc.) time, due to the conditioning. The value of an estimator can be averaged over all the rotation groups for the same period. The differential bias—the value relative to the average—can then be computed for each individual group to illustrate the relative effect of panel conditioning on the estimator.

Maintaining rotation groups of nearly equal size can be difficult, especially with units in business surveys. At the start of the rotation design, the sample units are usually selected and assigned to groups so that each has approximately the same weighted total of some key characteristic. However, over time, as units grow in size or drop out of existence, the groups may become unbalanced. This phenomenon can add variability to the estimates if composite estimation is used.

Issues of recall can develop when respondents are asked to provide data for different periods of time during the same interview. The *seam effect*, observed in some household panel longitudinal surveys, is caused by difficulty recalling the timing of events as respondents think back further in time. The opposite problem—early reporting bias—can affect economic surveys. Respondents asked to provide very recent sales data often report low, perhaps because their business accounts are incomplete and they underestimate their recent sales. The same respondents may provide more accurate reports for earlier periods of time because more data are available and the respondents' accounts are more complete.

Patrick J. Cantwell

See also Bureau of Labor Statistics (BLS); Composite Estimation; Current Population Survey (CPS); Longitudinal Studies; Panel; Panel Conditioning; Respondent Burden; Respondent Fatigue; Response Bias; Seam Effect; Statistics Canada; Survey Costs; U.S. Bureau of the Census

Further Readings

Cantwell, P. (1990). Variance formulae for composite estimators in one- and multi-level rotation designs. *Survey Methodology*, *16*(1), 153–163.

Cantwell, P., & Caldwell, C. (1998). Examining the revisions in monthly retail and wholesale trade surveys under a rotating panel design. *Journal of Official Statistics*, *14*(1), 47–59.

Kasprzyk, D., Kalton, G., Duncan, G., & Singh, M. P. (Eds.). (1989). *Panel surveys*. New York: Wiley.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business survey methods* (pp. 153–169). New York: Wiley.

U.S. Census Bureau. (2006, October). *Design and methodology* (Current Population Survey Tech. Paper 66). Retrieved March 4, 2008, from http://www.census.gov/prod/2006pubs/tp-66.pdf

Woodruff, R. S. (1963). The use of rotating samples in the Census Bureau's monthly surveys. *Journal of the American Statistical Association*, *58*, 454–467.

S

SALIENCY

Saliency refers to the degree to which a topic or event resonates with a prospective respondent or sample member. The more a topic or event resonates with a sample member, the more salient or important that topic or event tends to be in that person's life. Conversely, topics or events that resonate little or hold little importance for the sample member are said to have little saliency.

What are the implications of saliency for survey researchers? Saliency actually operates at two levels: the question level and the survey level.

On the question level, saliency refers to the importance of an event or action in a person's life. More important events or actions are better remembered than actions or events of low saliency. Consequently, saliency can affect the accuracy with which an event is remembered, which, in turn, can affect the accuracy of the response. More important or unusual events are generally remembered with greater accuracy than are common or frequent events. For example, most people can tell you, with little effort, their date of birth, the highest degree they have completed, what major illnesses they have suffered, or how many children they have. Similarly, given the significance of the event, many can tell you where they were when they heard the news of 9/11 or when John F. Kennedy was shot. Items of lesser importance, on the other hand, have lower saliency and are thus more difficult to remember. For example, recalling the number of times you have visited the grocery store in the past month or the number of movies you have seen in the past year can be difficult. On the other hand, while remembering how many movies you have seen in the past year is probably difficult, remembering the number of movies you have seen in the past week is probably not difficult. This illustrates an important point, the lower the saliency of an item, the shorter the reference period should be.

On the survey level, the saliency of the survey topic refers to the degree to which the subject matter of the survey resonates for the population being surveyed. If the questions being asked are of great interest to the average sample member, the survey is said to highly salient, whereas surveys where the subject being investigated is of little interest are said to have low saliency. Gaining cooperation or attaining a high response rate is made more difficult when the saliency is low, because sample members have little motivation to respond. On the other hand, when the central topic of a survey is one of great interest to those being surveyed, sample members are more likely to respond. For them, the burden of responding is compensated for by their interest in the topic. Thus, saliency is an important factor when thinking about response rates and the level of effort required to attain a certain response rate. For example, a questionnaire with high saliency and low respondent burden (e.g., takes minimal time to complete, is easy and straightforward to understand) will require much less effort to attain a high response rate than will a survey that has both low saliency and high respondent burden (i.e., takes a long time to complete, requires a great deal of record checking, or asks difficult or complex questions that require a great deal of thought). Surveys where the burden is high and the saliency is low often require respondent incentives to improve response rates.

Geraldine Mooney

See also Leverage-Saliency Theory; Reference Period; Telescoping

Further Readings

Groves, R. M., Singer, E. & Corning, A. (2000). Leverage-salience theory of survey participation. *Public Opinion Quarterly*, *64*, 299–308.

Sample

In survey research, a sample is a subset of elements drawn from a larger population. If all elements from the larger population are "sampled" for measurement, then a census is being conducted, not a sample survey.

In a broad context, survey researchers are interested in obtaining some type of information for some population, or universe, of interest. A sampling frame, that is, a frame that represents the population of interest, must be defined. The sampling frame may be identical to the population, or it may be only part of the population and is therefore subject to some undercoverage, or it may have an indirect relationship to the population (e. g., the population is males and the frame is telephone numbers). It is sometimes possible to obtain the desired information from the entire population through a census. Usually, however, for reasons of cost and time, survey researchers will only obtain information for part of it, referred to as a sample of the population. There may be several different samples selected, one for each stage of a multi-stage sample. For example, there may be a sample of counties, a sample of blocks within sampled counties, a sample of addresses within sampled blocks, and a sample of persons from sampled addresses.

A sample can be obtained in many different ways, as defined by the sample design. Survey researchers usually will want to have a probability sample, which ensures that all units in the frame have a known nonzero probability of selection, rather than a convenience sample or nonprobability sample, which do not sample respondents with known nonzero probabilities of selection.

Gary M. Shapiro

See also Census; Elements; Frame; Multi-Stage Sample; Nonprobability Sampling; Population; Population of Interest; Probability Sample; Representative Sample; Sample Design; Sampling; Sampling Frame; Undercoverage; Universe

Further Readings

Hansen, H. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: Wiley.
Särndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Sample Design

A sample design is the framework, or road map, that serves as the basis for the selection of a survey sample and affects many other important aspects of a survey as well. In a broad context, survey researchers are interested in obtaining some type of information through a survey for some population, or universe, of interest. One must define a sampling frame that represents the population of interest, from which a sample is to be drawn. The sampling frame may be identical to the population, or it may be only part of it and is therefore subject to some undercoverage, or it may have an indirect relationship to the population (e. g., the population is preschool children and the frame is a listing of preschools). The sample design provides the basic plan and methodology for selecting the sample.

A sample design can be simple or complex. For example, if the sampling frame consists of a list of every unit, together with its address, in the population of interest, and if a mail survey is to be conducted, then a simple list sampling would be appropriate; for example, the sample design is to have a sampling interval of 10 (select every 10th unit) from the list. The sample design must vary according to the nature of the frame and the type of survey to be conducted (the survey design). For example, a researcher may want to interview males through a telephone survey. In this case, the sample design might be a relatively simple one-stage sample of telephone numbers using random-digit dialing. One aspect of the sample design in this case is to determine whether all males in a household are to be interviewed and, if not, how to select a second-stage sample of males from each sampled telephone number that reaches a household.

The simplest type of sample design is purposive sampling, or convenience sampling. Usually, however, the survey researcher wants every unit in the frame to have a known probability of selection, so a more complex sample design is needed. In many situations, for purposes of efficiency and reducing costs, a multi-stage sample is desirable. For example, suppose a researcher wants to do face-to-face interviewing for a population consisting of African American women. In this case, the survey researcher might select an area probability sample, with the first stage of selection being a sample of counties, or primary sampling units. The next stage of selection might be a sample of blocks within the sampled counties, followed by a sample of housing units from the sampled blocks. The appropriate sampling frame at the first stage of selection for such a survey would be all counties in the United States. However, the researcher might decide to restrict the frame to only those counties with more than a particular percentage or number of African Americans. In this case, the survey researcher is introducing undercoverage into the sample design, which will likely result in some degree of bias in survey estimates. In deciding on the preferred sample design, the researcher must weigh the bias concerns against cost issues. If the researcher has a fixed budget, then sampling from all counties will result in higher variances or standard errors, but in lower bias, than will sampling from a restricted frame of counties.

The decisions that survey researchers make about their sample designs are among the most important ones that must be made to ensure the research is adequate for the information needs for which the survey is being conducted. Too often inadequate consideration is given in selecting a sample design, which, in turn, yields data that do not adequately meet the goals of the research. Often this happens because of the additional costs associated with conducting a survey using a more robust (and thus more appropriate) sample design.

Gary M. Shapiro

See also Adaptive Sampling; Area Probability Sample; Cluster Sample; Complex Sample Surveys; Convenience Sampling; Cross-Sectional Survey Design; Cutoff Sampling; Directory Sampling; Dual-Frame Sampling; Equal Probability of Selection; List Sampling; Multiple-Frame Sampling; Multi-Stage Sample; Network Sampling; Nonprobability Sampling; Probability of Selection; Probability Proportional to Size (PPS) Sampling; Probability Sample; Purposive Sample; Quota Sampling; Random-Digit Dialing (RDD); Random Sampling; Representative Sample; Sampling; Sampling Without Replacement; Sequential Sampling; Simple Random Sample; Snowball Sampling; Stratified Sampling; Systematic Sampling

Further Readings

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.
Hansen, H. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: Wiley.
Kish, L. (1995). *Survey sampling*. New York: Wiley.
Särndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Sample Management

In survey research, sample management refers to the efforts that must be made to coordinate the processing of sampled cases during a survey's field period so as to achieve the highest possible response rates within the finite budget that is allocated for data collection. This requires the coordination of whatever software is used to help manage the sample by the person(s) doing the managing.

The person or persons charged with sample management typically do this on a daily basis throughout the field period. For smaller surveys that involve a few hundred cases, this might be done by one person who may also have responsibility for other supervisory tasks. This person may manage the sample manually. In very large surveys with thousands of cases (including some large random-digit dialing surveys with millions of sampled telephone numbers), an entire team of supervisory personnel whose sole responsibility is sample management also will need to use specialized computer software to manage the sample.

Managing a survey sample differs greatly depending on whether or not the survey uses interviewers to gather the data. The following sections address sample management when the survey is interviewer-administrated and when it is self-administered.

Sample Management in Interviewer-Administered Surveys

In interviewer-assisted surveying, both for in-person and telephone surveys, sample management can be viewed as being composed of two primary responsibilities that often are in direct conflict with each other. One of these responsibilities is to allocate enough sample cases to keep interviewers productive and, in turn, to employ the right number of interviewers to process the sample so as to achieve the number of completed interviews that are required for the survey within the field period. The other responsibility is to activate "just enough" cases so that the field period ends with the highest possible response rates at the lowest possible cost.

Based on past experience, the sample manager of an interviewer-administered survey will have an informed expectation of approximately how many sample cases will need to be activated from the available sample (i.e., the *sampling pool*) during the survey field period to achieve the required sample size of completed interviews. Oftentimes the sample will be divided into sample replicates, each containing random subsets of cases. Once a replicate of cases has been activated, the entire replicate must be fully processed to allow each case to reach its proper final outcome.

If cases are not worked enough during the field period, response rates will suffer. Thus, the sample managers need to activate the last replicate of cases, leaving enough time in the field period to allow those cases to be processed fully. If not enough time is left to process this final set of cases, the most hard-to-reach respondents within those cases will not be contacted and interviewed. In contrast, if cases are contacted too frequently during the field period, for example, by having an overabundance of interviewers constantly making contact attempts, response rates also will suffer because respondents will become annoyed if the survey organization contacts them too frequently. Thus, the challenge for the sample managers is to have the right number of sample cases active at any one time and the right number of interviewers scheduled to process these cases.

To this end, the sample manager likely will need to track the progress that interviewers are making on the sample on a daily basis. The metrics that the manager will consider include the hourly or daily productivity of interviewers. Are they gaining completions more slowly than anticipated, faster than anticipated, or at about the rate anticipated? If productivity is too low, then more sample, more interviewers, or both, will be needed, unless the field period can be extended. Neither of these is appealing because both add costs to the survey budget. If productivity is higher than anticipated, then fewer interviewers will need to be scheduled or less sample will need to be activated during the remainder of the field. Making the correct decisions on these matters is important because costs can be saved, and this is highly attractive to survey organizations and their clients.

The success of any refusal conversion process that the survey may deploy is another metric that the sample manager will consider in determining whether the correct amount of sample has been activated or the correct number of interviewers has been scheduled, or both. If refusals are being converted at a rate (higher or lower) different from what was anticipated, the sample manager will need to make adjustments in allocating fresh cases for the duration of the field period.

There also are myriad personnel issues that survey sample managers face given the temporary and part-time nature of the work done by many survey interviewers. As such, interviewer turnover (churn) often is a problem that must be anticipated throughout a field period, especially for field periods that last more than a few weeks. Ultimately, the sample manager must be confident that there will be enough qualified interviewers to work the amount of sample that must be processed each day of the field period.

Depending on the size of the sample that must be processed within the field period, managing it in an interviewer-administered survey can be an extremely demanding and nerve-racking experience. Assuming that the quality of the final sample is important to the survey organization (although this is not always the case), even when using sophisticated software to help control the sample, idiosyncratic cases constantly will arise that need decisions made by a human sample manager on a case-by-case basis if high-quality sampling is to be achieved.

Sample Management in Self-Administered Surveys

In mail, Internet, and Web surveys, in which interviewers play no part in the data collection, managing the sample is much less complex. This is especially true if no follow-up mailings or other types of contacts with the nonresponding cases are planned to increase the survey's response rates.

In self-administered mail and Internet surveys, the sample cases typically are contacted initially, via mail or email, to inform them that they have been selected to participate in a survey. This contact may include the survey questionnaire, or it may mention that the questionnaire will be arriving within a few days, or it may direct the sampled respondent to a Web site where the questionnaire may be completed online. If the mode of contact and the mode of returning a questionnaire are via a postal delivery service, it is not uncommon for completed questionnaires to begin arriving back to the survey organization 2 days following the day of the original mailing. If the mode of contact is via the Internet, it is not unusual for some questionnaires to be completed within hours after the first emails have been sent.

The sample manager(s) will need to track incoming mailed-back questionnaires or questionnaires completed on a Web site on a daily basis. In the case of questionnaires that are mailed back, there is a labor-intensive process that must be staffed to receive, open, screen for completeness, and log in the questionnaires. The managers will have expectations based on prior experience and the specific design features of the survey (e.g., the use of incentives to stimulate response) about what level of response is anticipated. The vast majority of replies to an initial mailing will come back within 2 weeks of the original mailing. If the survey has budgeted a follow-up mailing (or multiple follow-up mailings) to raise the response rate, then the sample managers will need to work with the staff that implements the follow-up mailing with enough lead time so that the second mailing can go out on a timely basis. In most survey designs, the researchers will know which specific cases already have responded prior to the second mailing and thus will be able to avoid the need to do a full follow-up mailing to all original sampled cases. However, it is impossible to avoid mailing a follow-up to some of the cases who already responded to the original mailing because (a) some of their returns will not arrive by the time of the follow-up mailing, and (b) some will never be received because of mail delivery or Internet problems.

Sample Management in Mixed-Mode Surveys

As response rates have dropped during the past two decades, the need for researchers to implement mixed-mode surveys, which often can achieve higher response rates at cost-favorable levels, has become more commonplace. For example, using an address-based sampling frame would allow a researcher to implement a design that begins the field period with a mail survey, as well as offering an Internet mode for completing the questionnaire and other survey task(s). Then, nonresponders for whom the researchers have a telephone number matched to their address would be followed up using telephone interviewers. Finally, in-person interviewers would be sent during the last stage of the field period to all addresses that have not responded to previous modes of contact. Managing a mixed-mode sample design such as this is much more complex than managing a sample for a survey that uses a single mode of data collection. In the case of the mixed-mode approach, with multiple channels allowed for data to be gathered and to be returned to the survey organization, it is paramount that the sample managers have the proper software systems in place to capture and safely store the incoming data and to accurately track the status of every case on a daily basis that remains active. If this does not happen, the response rates for the survey will suffer in ways that could have been avoided had a better sample management system been deployed.

Paul J. Lavrakas

See also Address-Based Sampling; Case; Case-Control Study; Contact Rate; Control Sheet; Face-to-Face Interviewing; Field Period; Hit Rate; Internet Surveys; Mail Survey; Mixed Mode; Refusal Conversion; Response Rates; Sample; Sample Replicates; Sampling Pool; Supervisor; Supervisor-to-Interviewer Ratio; Survey Costs; Telephone Surveys; Total Design Method (TDM); Web Survey

Further Readings

Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z., Martin, J., Nicholls, W. L., et al. (1998). *Computer assisted survey information collection*. New York: Wiley.

Hansen, S. E. (2007). CATI sample management. In J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw,

L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 340–358). New York: Wiley.
Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Sample Precinct

Sample precinct refers to a sampling unit used in data collection and analysis on Election Day. The term is most commonly associated with media exit polling and election night projections. For voting purposes, jurisdictions such as counties or townships are divided into precincts based on geography. A precinct is the smallest voting unit in a jurisdiction. Voters are assigned to a precinct, which typically has one polling place where voters can go to cast their ballots. Data are collected from sample precincts, which are used to form estimates of the vote and voter opinions and characteristics.

For Election Day analysis, a sample of precincts is selected because it would be too costly and time consuming to collect data from all precincts in a jurisdiction (e.g., a state or city). Prior to the election, a representative sample of precincts is selected for a given jurisdiction. A listing of all precincts in the jurisdiction is compiled, and a probability sample of precincts is selected. Typically, stratified sampling is used to increase the precision of the estimates. Precincts can be stratified using such variables as historical voting patterns, geography, and race/ethnicity.

In practice, sample precincts are used for two purposes. The first is to provide data to project the outcome of the election using actual votes from the sample precincts. Workers, sometimes called *stringers*, are sent to the sample precincts on Election Day. As soon as possible after the polling place closes, the stringer's job is to get the actual vote totals from the polling place official and call these results into a central location where they are tabulated and fed into computerized statistical models for analysis. Sometimes it is also possible to get the sample precinct votes via the Internet or by phone. The sample precinct models can take various forms, including ratio and regression estimators. For statewide elections, a typical sample size for this purpose is about 80 precincts and usually varies from 15 to 100 depending on the characteristics of the state, the newsworthiness of the electoral contest, and how close the election is expected to be.

The second use of precinct samples is for conducting Election Day exit polls, such as those conducted for the National Election Pool by Edison Media Research each major election year since 2004. The exit poll precinct sample is usually a smaller subset of the sample used to collect the actual vote tallies. A typical statewide exit poll will have about 30 sample precincts and typically ranges from 15 to 60. Exit poll interviewers are sent to the sample precincts to interview voters as they exit the polling place. Data are collected on voters' demographics and opinions and how they voted. These data are called into a central location three times during the day and tabulated. The demographic and opinion data are cross-tabulated with the vote and are used to analyze why voters voted the way they did. These data help explain why candidates won or lost and what issues were important to voters. The data from the exit poll sample precincts are also used to help project the election outcomes.

Daniel M. Merkle

See also Election Night Projections; Election Polls; Exit Polls; National Election Pool (NEP); Stratified Sampling

Further Readings

Merkle, D. M., & Edelman, M. (2000). A review of the 1996 voter news service exit polls from a total survey error perspective. In P. J. Lavrakas & M. Traugott (Eds.), *Election polls, the news media, and democracy*. New York: Chatham House.

Sample Replicates

A sample replicate is a random subset of the entire available sample (i.e., *sampling pool*) that has been drawn for a particular survey. Sample replicates help survey managers coordinate the progress that is made on data collection during the survey's field period. Sample replicates often are made up of a randomly assigned 1,000 of the sampled elements, although sometimes replicates may be as small in size as 100.

The value of structuring a survey sampling pool into replicates is that it is not possible to know in advance exactly how many of the telephone numbers, emails, or street addresses in a sampling pool actually will need to be released and processed in order to achieve the final sample size of completed interviews a given survey needs. Sometimes survey organizations make better progress than expected in processing their samples, whereas other times they make worse progress. Releasing the sample in these replicates prevents releasing more sampled elements than are needed, which in turn helps control survey costs.

A rule of thumb that is used by most survey organizations is that once a replicate has been released, all the elements in that replicate must be fully processed according to whatever contact rules are being applied. Doing otherwise would lead to a distortion of the final sample with the last-to-be-activated replicates having data collection take place only from the easiest-to-reach respondents in these replicates, whereas data collection would not be successful from the harder-to-reach respondents in the last-to-be-activated replicates. This would conflict with what will have taken place with replicates that were released earlier in the field period, where all possible completed interviews were attained regardless of how difficult it was to reach the respondents.

Organizing a sampling pool into replicates is straightforward and essentially occurs in one of two ways. First, a "large enough" sampling pool should be drawn from the sampling frame so that there are more elements drawn than would ever be expected to be needed to complete the survey. Then these elements are cleaned (e.g., duplicates are eliminated) and then randomly ordered and then divided (segmented) into replicates of whatever size is desired, for example, of size 1,000. This first approach works best when the sampling frame is not already in a random order. In cases where the sampling frame can be accessed randomly, then replicates can be created directly as elements are drawn from the frame. If the replicates are to be 1,000 in size, then under this second approach the first 1,000 elements drawn make up the first replicate, the second 1,000 drawn make up the second replicate, and so on. When using this second approach, safeguards must be used to make certain there will not be duplicate elements in different replicates.

Supervisory personnel who are responsible for controlling the data collection during the field period will generally start by releasing enough replicates to carry the interviewing staff through the first one third or one half of the field period. Based on the productivity that is observed in this early stage (e.g., looking at the pattern of the final dispositions of the elements that have been processed to date), the sample managers will calculate whether the initial estimates of how much sample will be needed look to be on track, too low, or too high. If on track, nothing will be done about the amount of sample that will be processed. If too high, some of the replicates that were expected to be needed will be held back. If too low, more replicates will need to be released before the field period ends.

Managing the release and processing of sample replicates is very important for the end quality of a survey and its costs. Releasing too many replicates may lead to an erosion of the survey's response rates because there may not be enough staff or time to fully process the sample during the remainder of the field period. Releasing too few replicates will lead to staff not having enough sampled cases to work, and thus staff will not be able to operate at peak efficiency.

Paul J. Lavrakas

See also Calling Rules; Elements; Field Period; Final Dispositions; Sample; Sample Management; Sample Size; Sampling Frame; Sampling Pool; Survey Costs

Further Readings

Hansen, S. E. (2007). CATI sample management systems. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.), *Advances in telephone survey methodology* (pp. 340–358). New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Sample Size

The sample size of a survey most typically refers to the number of units that were chosen from which data were gathered. However, sample size can be defined in various ways. There is the *designated sample size*, which is the number of sample units selected for contact or data collection. There is also the *final sample size*, which is the number of completed interviews or units for which data are actually collected. The final sample size may be much smaller than the designated sample size if there is considerable nonresponse, ineligibility, or both. Not all the units in the designated sample may need to be processed if productivity in completing interviews is much higher than anticipated to achieve the final sample size. However, this assumes that units have been activated from the designated sample in a random fashion. Survey researchers may also be interested in the sample size for subgroups of the full sample.

In planning to conduct a survey, the survey researcher must decide on the sample design. Sample size is one aspect of the sample design. It is inversely related to the variance, or standard error of survey estimates, and is a determining factor in the cost of the survey. In the simplest situation, the variance is a direct function of the sample size. For example, if a researcher is taking a simple random sample and is interested in estimating a proportion p, then

$$\text{Variance} = (p)(1-p)/n,$$

where n is the sample size. More generally,

$$\text{Variance} = f(p)(1-p)/n,$$

where f is the design effect, which reflects the effect of the sample design and weighting on the variance.

In the planning effort for a more complex survey, it is preferable to not focus directly on sample size. It is best to either (a) set a budget and determine the sample size and sample design that minimize the variance for the available budget, or (b) set a desired variance or required reliability, possibly using statistical power analysis, and then determine the sample size and sample design that minimize costs while achieving the desired variance. Sample size does not solely determine either the variance or the cost of a survey and thus is not generally, by itself, a meaningful planning criterion. More useful is *effective sample size*, which adjusts the sample size for the sample design, weighting, and other aspects of the survey operation.

Even better than fixing the budget and minimizing the variance is to minimize the mean square error (MSE), the researcher can set a desired MSE and minimize the cost to obtain it. MSE is defined as the sum of the variance and the bias squared and thus accounts for more than just the variance. One sample design option may have a larger sample size and a lower variance than a second option but have a larger MSE, and thus it would be a poor choice.

A common misconception is that the needed sample size is a function of the size of the population of interest, or universe. For example, people often think that to achieve a given precision, a much larger sample size is required for a sample of the entire United States than of a large city. This is not generally true. However, if a survey researcher is considering a sample that is a substantial proportion of the population of interest, he or she might apply a finite population adjustment that reduces the variance. Thus, in the case of a sample that is large relative to the population, the needed sample size is reduced. When the researcher is interested in a superpopulation, which is much larger than the actual sampling frame, or is interested in an analytic survey, then the finite population adjustment should not be applied.

Frequently, the need is to determine the sample size for a given sample design that will produce "sufficiently reliable" estimates. There are a couple of ways to address this. One way is to estimate the standard error, or variance, that would be obtained for various sample size choices. The needed standard error, the needed *coefficient of variation*, or the needed confidence interval can be determined. The coefficient of variation is the standard error divided by the estimate and is sometimes more useful to consider than the standard error itself. A rule of thumb that is sometimes used for what is sufficiently reliable is that the coefficient of variation be no more than .10. Setting criteria on the desired size for a 90% or 95% confidence interval is often a useful method for determining what sample size is needed. Estimation of standard errors can be difficult with a complex sample design or if little is known about the population distribution for a characteristic of interest.

Another approach that often is used to determine sample size for a sufficiently reliable estimate is statistical power analysis. This considers both so-called Type I error or alpha (probability of rejecting a true null hypothesis) and Type II error (the probability of failing to reject a false null hypothesis) in determining the needed sample size. The researcher needs to specify a hypothesis test, an effect size, and an acceptable Type II error to be able to perform this calculation.

Gary M. Shapiro

See also Confidence Interval; Design Effects (*deff*); Effective Sample Size; Finite Population Correction (fpc) Factor; Mean Square Error; Population of Interest; *p*-Value; Sample Design; Sampling Pool; Statistical Power; Superpopulation; Type I Error; Type II Error; Variance

Further Readings

Bell, J. (1998). Big is not necessarily beautiful in survey design: Measurement error and the APU Science Survey. *Statistician*, *40*, 291–300.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.
Lenth, R. V. (2001). Some practical guidelines for effective sample-size determination. *American Statistician*, *55*, 187–193.
Milton, S. (1986). A sample size formula for multiple regression studies. *Public Opinion Quarterly*, *50*, 112–118.
Rust, K., Graubard, B., Fuller, W. A., Stokes, S. L., & Kott, P. S. (2006). Finite population correction factors [Panel discussion]. *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. Retrieved May 23, 2008, from http://www.cos.gmu.edu/~wss/wss070328paper.pdf

Sampling

Sampling is the selection of a given number of units of analysis (people, households, firms, etc.), called cases, from a population of interest. Generally, the sample size (n) is chosen in order to reproduce, on a small scale, some characteristics of the whole population (N).

Sampling is a key issue in social research designs. The advantages of sampling are evident: feasibility of the research, lower costs, economy of time, and better organization of the work. But there is an important problem to deal with: that is, sampling error, because a sample is a model of reality (like a map, a doll, or an MP3) and not the reality itself. The sampling error measures this inevitable distance of the model from reality. Obviously, the less it is, the more the estimates are close to reality. Unfortunately, in some cases, the sampling error is unknowable.

There are two main families of sampling methods: probability (random) sampling and nonprobability sampling, respectively typical of (but not exclusive to) quantitative and qualitative research.

Probability sampling, definitively codified in the 1930s by the Polish statistician Jerzy Neyman, is characterized by the condition that all units of the population have an (theoretically) equal, calculable (i.e., known), nonzero probability of being included in the sample.

Probabilistic samples are considered representative of reality: What can be said about the sample can be extended to the reality of what is sampled by statistical inference. Another advantage is that the sampling error, which is a crucial datum to assess the validity of the sample, is calculable: This is possible only for probability samples. The main problem, however, is that researchers need the complete list of the target population (i.e., the sample frame), though sometimes the exact number of the population is sufficient, to extract the sample, and often this is impossible to obtain (e.g., when a researcher wants to study the audience of a movie).

There are several types of probability sampling. The most common are simple, systematic, and stratified random sampling. Other types of probability samples are multi-stage, cluster, multi-phase, and spatial sampling.

In most cases, the size of a probability sample is determined by the following formula:

$$n = \frac{z^2 pqN}{E^2(N-1) + z^2 pq},$$

where z refers to the confidence level of the estimate (usually fixed at 1.96, corresponding to a 95% confidence level), pq is the variance (that is unknown and then fixed at its maximum value: 0.25), N is the size of the population, E is the sampling error (often ≤ 0.04).

Nonprobability samples are generally purposive or theory driven. This means they are gathered following a criterion the researcher believes to be satisfying to obtain typological representativeness. This latter is achieved, when the researcher has sufficient members of all the main categories of interest to be able to describe with confidence their patterned similarities and differences.

Being purposive, nonprobability samples are rather heterogeneous. Up to 16 different qualitative sampling strategies have been listed for choosing a nonprobability sample. It is almost impossible to give an exhaustive list, because they are continuously open to integrations and new solutions. However, quota, snowball, purposive, theoretical, and accidental sampling are among the most common types of nonprobability sampling techniques.

The main problem with nonprobability samples is that the researcher has only loose criteria for assessing their validity: The sampling error is unknowable, so the researchers cannot say whether the results are representative or not, and the risk of nonsampling errors is large.

The big issue with sampling remains representativeness (i.e., external validity). Are probability samples really representative? The answer to this question is not trivial. In fact, probability samples cannot guarantee representativeness, for at least four reasons:

1. Survey researchers cannot say whether a sample is indeed representative or not, because they generally sample precisely to find out something about an unknown reality. This is the so-called *sampling paradox*.

2. To prove or correct (in case of post-stratification) the representativeness of a sample, the estimates are often compared to census data. In this case, the researcher must take into account two further problems: (a) Census data may be too old, and (b) they could represent a benchmark only with respect to certain variables, so the best thing the researcher can obtain is a sample that is representative only with respect to those limited variables (mainly demographics).

3. The researcher must take into account nonsampling errors, trying to minimize them (e.g., through weighting). There are four major types of nonsampling errors: coverage errors (e.g., when the list of the population is incomplete), measurement errors (due to bad questionnaires); nonresponse errors (associated with refusals, noncontacts, movers, illiteracy, language barriers, and missing data); and processing errors (coding or inputting errors). These errors are often quite difficult to know and control.

4. Nonprobability samples may be representative by chance (e.g., many quota samples prove to be representative a posteriori).

These are the reasons why, on one hand, nonprobability samples are used even in important surveys, and, on the other, a hybrid direction is gradually getting a footing in the social research community, as the success of mixed strategies like respondent-driven sampling shows.

A recent frontier in sampling is the alliance with the new technologies. Though rather promising, however, Internet sampling, cell-phone sampling, and others still have to deal with many problems. For example, the number of Internet users is significantly lower among older people. For this reason, some sort of adjustment or sampling mix often must be considered, and even then the results may not be representative.

Alberto Trobia

See also Case; Cluster Sample; Element; External Validity; Multi-Stage Sample; Nonprobability Sampling; Probability Sample; Purposive Sample; Quota Sampling; Representative Sample; Respondent-Driven Sampling (RDS); Sample Design; Sample Size; Sampling Error; Simple Random Sample; Snowball Sampling; Stratified Sampling; Systematic Sampling

Further Readings

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Sudman, S., & Blair, E. (1999). Sampling in the twenty-first century. *Journal of the Academy of Marketing Science*, *27*(2), 269–277.

Sampling Bias

Sampling bias occurs when a sample statistic does not accurately reflect the true value of the parameter in the target population, for example, when the average age for the sample observations does not accurately reflect the true average of the members of the target population. Typically, sampling bias focuses on one of two types of statistics: averages and ratios. The sources of sampling bias for these two types of statistics derive from different sources; consequently, these will be treated separately in this entry.

Sampling Bias for Averages

For survey researchers, sampling biases for averages derive from three sources: (1) imperfect sampling frames, (2) nonresponse bias, and (3) measurement error. Mathematical statisticians may also consider biases due to sources such as using the sample size (n) instead of $n-1$, or using a sample statistic (e.g., s^2) to estimate a population parameter (e.g., σ^2), but these tend to be of academic interest and of less interest to practical research concerns; therefore, these will not be considered in this entry.

Imperfect sampling frames occur frequently in research and can be classified into four general categories: (1) frames where elements are missing, (2) frames where elements cluster, (3) frames that include foreign elements, and (4) frames with duplicate listings of elements. For example, household telephone surveys using random-digit dialing samples for landline telephones exclude households that are cell phone only (missing elements), include telephone numbers in some dwelling units that include more than a single household (element cluster), include some telephone numbers that are dedicated solely to fax machines (foreign elements), and include some households with more than a single landline telephone number (duplicate listings). All of these can cause sampling bias if they are not taken into account in the analysis stage.

Sampling bias due to nonresponse results from missing elements that should have been included in the sample but were not; these can be classified as noncontact (missing), unable to answer, or refusal. Noncontacts are those elements that are selected into the sample but cannot be located or for which no contact can be made; whereas the unable-to-answer either do not have the necessary information or the required health or skills to provide the answer, refusals are elements that, once located, decline to participate. Each will contribute to sampling bias if the sampled nonresponding elements differ from the sampled elements from which data are gathered. This sampling bias for averages can be characterized as follows:

$$\bar{Y}_{responders} - \bar{Y}_{population} = \frac{n_{nonresponders}}{n_{sample}}\left(\bar{Y}_{responders} - \bar{Y}_{nonresponders}\right).$$

As can be seen, when the number of nonresponse represents a small proportion of the total sample, or when there is a small difference between those who respond and those who do not, the resulting sampling bias will be small or modest. If it is possible to place bounds on the averages (such as with probabilities and proportions), researchers can quantify the possible range of sampling bias.

The third source of sampling bias for averages results from measurement error—that is, when what is measured among the sample elements differs from what researchers actually wished to measure for the target population. Measurement error, however, can be divided into *random error* and *consistent bias*; only consistent bias results in sampling bias, as random error will appear as sampling variance. These components of measurement error can be represented as

$$y_{ij} = \mu_i + \beta_i + \varepsilon_{ij},$$

where y_{ij} represents the observed values of some variable y on j repeated observations of individual i, μ_i represents the true value of what the researchers wish to measure for individual i, β_i represents the consistent bias in individual i's response, and ε_{ij} represents the random error associated with observation j for individual i. In sample surveys, consistent measurement bias can occur for a number of reasons, such as the questions focus on issues that are subject to a social desirability bias (e.g., illegal drug use).

Sampling Bias for Ratios

Unlike the estimation of sample averages, ratio estimators computed from samples are biased. In the 1950s, Herman Hartley and A. Ross showed that the absolute amount of bias in a ratio estimator is small relative to its standard deviation if the coefficient of variation (i.e., the ratio of the standard deviation to the average) of the ratio's denominator variable is small. In general terms, the amount of bias in ratio estimates will be small if the sample size is large, the sampling fraction is large, the mean of the denominator variable is large, the variance of the denominator variable is small, or if the correlation between numerator and denominator variables in the ratio is close to positive 1.0.

Allan L. McCutcheon

See also Measurement Error; Nonresponse Bias; Parameter; Random Error; Sampling Frame; Social Desirability; Statistic; Target Population; True Value

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Hartley, H. O., & Ross, A. (1954). Unbiased ratio estimators. *Nature*, *174*, 270–271.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Sampling Error

Sampling error consists of two components: *sampling variance* and *sampling bias*. Sometimes overall sampling error is referred to as *sampling mean squared* *error* (MSE), which can be decomposed as in the following formula:

$$\begin{aligned} MSE(p) &= E(p - P) \\ &= E[(p - p') + (p' - P)] \\ &= \mathrm{Var}(y) + \mathrm{Bias}^2, \end{aligned} \quad (1)$$

where P is the true population value, p is the measured sample estimate, and p' is the hypothetical mean value of realizations of p averaged across all possible replications of the sampling process producing p.

Sampling variance is the part that can be controlled by sample design factors such as sample size, clustering strategies, stratification, and estimation procedures. It is the error that reflects the extent to which repeated replications of the sampling process result in different estimates. Sampling variance is the random component of sampling error since it results from "luck of the draw" and the specific population elements that are included in each sample. The presence of sampling bias, on the other hand, indicates that there is a systematic error that is present no matter how many times the sample is drawn.

Using an analogy with archery, when all the arrows are clustered tightly around the bull's-eye we say we have low variance and low bias. At the other extreme, if the arrows are widely scattered over the target and the midpoint of the arrows is off-center, we say we have high variance and high bias. In-between situations occur when the arrows are tightly clustered but far off-target, which is a situation of low variance and high bias. Finally, if the arrows are on-target but widely scattered, we have high variance coupled with low bias.

Efficient samples that result in estimates that are close to each other and to the corresponding population value are said to have low sampling variance, low sampling bias, and low overall sampling error. At the other extreme, samples that yield estimates that fluctuate widely and vary significantly from the corresponding population values are said to have high sampling variance, high sampling bias, and high overall sampling error. By the same token, samples can have average level sampling error by achieving high levels of sampling variance combined with low levels of sampling bias, or vice versa. (In this discussion it is assumed, for the sake of explanation, that the samples are drawn repeatedly and measurements are made for each drawn sample. In practice, of course, this is not feasible, but the repeated measurement scenario serves as a heuristic tool to help explain the concept of sampling variance.)

Sampling Variance

Sampling variance can be measured, and there exist extensive theory and software that allow for its calculation. All random samples are subject to sampling variance that is due to the fact that not all elements in the population are included in the sample and each random sample will consist of a different combination of population elements and thus will produce different estimates. The extent to which these estimates differ across all possible estimates is known as sampling variance. Inefficient designs that employ no or weak stratification will result in samples and estimates that fluctuate widely. On the other hand, if the design incorporates effective stratification strategies and minimal clustering, it is possible to have samples whose estimates are very similar, thereby generating low variance between estimates, thus achieving high levels of sampling precision.

The main design feature that influences sampling variance is sample size. This can be seen readily from the following formula for the sampling variance to estimate a proportion based on a simple random sample design:

$$var(p) = pq/n, \quad (2)$$

where p is the sample estimate of the population proportion, $q = 1 - p$, and n is the sample size. (Formula 2 and subsequent formulae are relevant for proportions. Similar formulae are available for other statistics such as means, but they are more complicated.)

It can be easily seen from this formula that as n increases, the variance decreases in direct and inverse proportion. Because sampling variance is usually measured in terms of the confidence interval and standard error (which is the square root of the sampling variance), we usually refer to the impact of an increase in sample size in terms of the square root of that increase. Thus, to double the sampling precision, that is, reduce the sampling variance by 50%, we would have to increase the sample size by a factor of 4.

Sampling variance, or its inverse, sampling precision, is usually reported in terms of the standard error, confidence interval, or more popularly, the margin of error. Under a simple random sample design, the mathematical formula for the standard error (3) and the 95% confidence interval for a proportion p (4) are

$$\mathrm{se}(p) = [\mathrm{var}(p)]^2 \quad (3)$$

$$ci(p) = [p - 1.96\mathrm{se}(p), p + 1.96\mathrm{se}(p)]. \quad (4)$$

The margin of sampling error is equal to half the length of the confidence interval as defined in Formula 4. For example, a proportion of 50% from a sample size of 1,000 would have a margin of error of "plus or minus 3%," meaning that if we were to draw 100 simple random samples of approximate size of 1,000, for about 95 of the samples, the sample value would differ by no more than 3 percentage points in either direction from the true population value.

In general, the main drivers of sampling variance are stratification and clustering. Stratification usually results in a lower sampling variance because the number of possible samples is reduced in comparison with an unrestricted simple random sample. Not only is the number of possible samples reduced, but potential outliers are eliminated. For example, suppose we wanted to sample households in the United States. An unrestricted random sample might contain households that are all located in the Northeast—the probability is not high, but it is not zero. However, if we stratify by region, then we reduce the probability of such a skewed sample to zero. To the extent that the variable of interest is related to our stratification variables, in this case geography, stratification will reduce the overall sampling variance. In setting up the design, therefore, it is important to strive to define strata that are relatively homogeneous with respect to the variables of interest.

Clustering, on the other hand, works in a very different way. Clustering plays a role in sample designs that are used for surveys in which the data are collected in person, for example, via household visits. Clustering is used to control field costs, especially those related to travel, which often represent a significant portion of the overall survey budget. However, this results in fewer degrees of freedom in the sense that the sample now focuses on a smaller number of sampling units, that is, the first-stage clusters, often referred to as primary sampling units. For example, selecting an unclustered sample of 1,000 households throughout the United States would mean that the households could be located anywhere in the country and, of course, this would result in large travel costs. Restricting the sample first to 100 clusters (e.g., counties), and then taking 10 households within each cluster, reduces the travel costs but reduces our ability to spread the sample effectively over the entire country. This reduction in efficiency is further exacerbated by the fact that within each cluster, usually a geographically contiguous area, households tend to be more alike than households across these units. This phenomenon, called *intraclass homogeneity*, tends to drive up the sampling variance because efficiency is lost and the original sample of 1,000 might, in effect, have only the impact of 100 if, in the extreme case, the clusters are perfectly homogeneous.

Thus, in summary, with respect to sample design optimization, stratification is beneficial in that it reduces sampling variance, whereas clustering is to be avoided when possible or at least minimized as its effect is to increase sampling variance. Usually the effect of clustering is more marked than that of stratification. In many situations, though, clustering is necessary for cost reasons; thus, the best clustered design strategy involves finding a compromise between the cost savings and the penalty to be paid in terms of lower precision.

Another important factor that influences sampling variance is weighting. Weighting refers to adjustment factors that account for design deviations such as unequal probabilities of selection, variable nonresponse rates, and the unavoidable introduction of bias at various steps in the survey process that are corrected for through a process called *post-stratification*. The effect of weighting is to increase the sampling variance, and the extent of this increase is proportional to the variance among the weights.

A useful concept that quantifies and summarizes the impact of stratification, clustering, and weighting on sampling variance is the *design effect*, usually abbreviated as *deff*. It is the ratio of the true sampling variance taking into account all the complexities of the design to the variance that would have been achieved if the sample had been drawn using a simple random sample, incorporating no stratification, clustering, or weighting. A value of 1.00 indicates that the complexity of the design had no measurable impact on the sampling variance. Values less than 1.00 are rare; values larger than 5.00 are generally considered to be high.

The design effect is closely related to *rho* (ρ), the intraclass correlation, mentioned previously. The following formula shows the relationship between the two:

$$deff = 1 + (b - 1)\rho, \quad (5)$$

where *deff* is the design effect, ρ is the intraclass correlation, and b is the average cluster size.

The correct calculation of sampling variance, incorporating all the complexities of the design, is not straightforward. However, there is extensive software currently available that uses either the empirical bootstrap replication approach or the more theoretically based Taylor Series expansion. These systems typically allow for many types of stratification, clustering, and weighting although the onus is always on the user or the data producer to ensure that relevant information, such as the stratum identifier, cluster identifier, and weight, are present in the data set.

Sampling Bias

This component of sampling error results from a systematic source that causes the sampling estimates, averaged over all realizations of the sample, to differ consistently from their true target population values. Whereas sampling variance can be controlled through design features such as sample size, stratification, and clustering, we need to turn to other methods to control and reduce bias as much as possible.

Sampling bias can only be measured if we have access to corresponding population values. Of course, the skeptic will point out that if such information were available, there would be little point in drawing a sample and implementing a survey. However, there are situations in which we can approximate sampling bias by comparing underlying information such as basic demographics for the sample with corresponding data from another, more reliable, source (e.g., census or large national survey) to identify areas in the data space for which the sample might be underrepresented or overrepresented.

One major source of sampling bias is *frame coverage*; that is, the frame from which the sample is drawn is defective in that it fails to include all elements in the population. This is a serious error because it cannot be detected, and in some cases its impact cannot even be measured. This issue is referred to as *undercoverage* because the frame is missing elements that it should contain. The opposite phenomenon, *overcoverage*, is less serious. Overcoverage occurs when the frame includes foreign elements, that is, elements that do not belong to the target population. However, these elements, if sampled, can be identified during the field operation and excluded from further processing. A third potential source of frame bias is *duplication*. If certain elements appear several times on the frame, their probabilities of selection are higher and thus they might be overrepresented in the sample. Furthermore, it is not always known how many times the elements occur on the frame, in which case it is impossible to ascertain the extent of the problem and thus the size of the bias.

Sampling bias can also occur as a result of flaws in the sample selection process, errors in the sample implementation, and programming missteps during the sample processing stage. An example of bias occurring during sample selection would be a systematic sample of every fifth unit when, in fact, there is a repeating pattern in the list and every fifth unit belongs to a special group. An example of how sampling bias can occur during the sample implementation process is the method interviewers use to visit households in the field. Field instructions might indicate "every 10th household," and the interviewer might instead elect to visit households that appear more likely to generate an interview. This could, and often does, lead to sampling bias. Finally, sampling or estimation bias can occur during the sample processing stage, for example, by incorrect calculation of the weighting adjustment factors, giving excessive importance to certain subpopulations.

One severe challenge faced by all survey practitioners is how to measure bias. Whereas the estimation of sampling variance emanates from statistical theory (see Formula 2, presented earlier), the only way to measure sampling bias is to compare the resulting empirical value with the true target population value. Of course, this is problematic because we seldom possess the population value and thus must use indirect methods to estimate bias. One approach uses data from other sources, such as the census or large national samples, as surrogates for the population being sampled. The problem with this strategy is that even the census is subject to error, in terms of both variance and bias.

It was pointed out previously that weighting tends to increase sampling variance and reduce precision. The reason weighting is implemented in survey research, in spite of its negative effect on variance, is that in many cases it can be used to reduce bias by bringing the sampling distributions more in line with known population distributions. For example, it is often possible to weight to basic census distributions by gender and age, even for minor geographical subdivisions such as tracts. To take a hypothetical example, suppose the sample distribution by gender turns out to be 40% male, 60% female, a not uncommon result in a typical random-digit dialing telephone survey. Furthermore, assume that the corresponding census numbers are close to 50:50. Weighting would assign relative adjustment factors of 50/40 to males and 50/60 to females, thus removing the possible bias due to an overrepresentation of females in the sample.

Challenges

Overall sampling error needs to be viewed in terms of a combination of sampling variance and sample bias. The ultimate goal is to minimize the mean squared error. Survey researchers know how to measure sampling variance, and they have a good handle on how it can be reduced. Sampling bias represents more of a challenge as it is often difficult to measure and even if it is measurable, bias reduction is often expensive and problematic to achieve.

It is illustrative to discuss surveys that are based on nonprobability judgment, quota, or convenience samples—that is, samples that are not based on probability-based design. One currently prominent example is the Internet-based panel, which consists of members who choose (self-select) to belong to these panels. That is, the panel members are not selected randomly and then invited to join the panel, but rather, the members themselves decide to join the panels, hence the term *opt-in* populations. This means that the underlying frame suffers from undercoverage and many potential types of bias, only some of which are known. These samples might be appropriate for certain studies (e.g., focus groups), in which generalizing with confidence to the population is not an absolute prerequisite. But, in general, these surveys fall short of required methodological rigor on two counts. In the first place, the probabilities of selection are usually unknown and often unknowable, thus precluding any chance of calculating sampling variance. Second, these surveys suffer from coverage and selection bias issues that, in many cases, are not even measurable.

With the advent of relevant software, surveys now regularly produce large-scale sampling variance results showing not only standard errors and confidence intervals but also design effects and measures of intraclass correlation. The results typically are presented for the entire sample and also for important subpopulations that are relevant for data users. These are useful not only to shed light on the quality of the data but also to inform future sample designs. The choice of estimates and subpopulations for which to publish sampling errors is not simple, and some researchers have developed "generalized variance functions" that allow users to estimate their own sampling errors based on the type of variables in question, the sample size, and level of clustering. However, these results are usually limited to sampling variance, and much less is calculated, produced, and disseminated with respect to sampling bias. This is due largely to the difficulty of calculating these measures and to the challenge of separating sampling bias from other sources of bias, such as nonresponse bias and response bias.

Karol Krotki

See also Bootstrapping; Clustering; Coverage Error; Design Effects (*deff*); Duplication; Intracluster Homogeneity; Margin of Error (MOE); Mean Square Error; Nonprobability Sampling; Nonresponse Bias; Overcoverage; Post-Stratification; Probability of Selection; Response Bias; ρ (Rho); Sample Size; Sampling Bias; Sampling Frame; Sampling Variance; Self-Selection Bias; Simple Random Sample; Strata; Stratified Sampling; Systematic Sampling; Taylor Series Linearization; Undercoverage; Weighting

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.
Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Sampling Fraction

A sampling fraction, denoted f, is the proportion of a universe that is selected for a sample. The sampling fraction is important for survey estimation because in sampling without replacement, the sample variance is reduced by a factor of $(1-f)$, called the *finite population correction* or *adjustment*.

In a simple survey design, if a sample of n is selected with equal probability from a universe of N, then the sampling fraction is defined as $f = n/N$. In this case, the sampling fraction is equal to the probability of selection. In the case of systematic sampling, $f = 1/I$ where I is the sampling interval.

The sampling fraction can also be computed for stratified and multi-stage samples. In a stratified (single-stage) sample, the sampling fraction, f_h is computed separately for each of the h strata. For a stratified sample $f_h = n_h/N_h$, where n_h is the sample size for stratum h and N_h is number of units (in the universe of N) that belong to stratum h. Because many samples use stratification to facilitate oversampling, the probabilities of selection may differ among strata, in which case the f_h values will not be equal.

For multi-stage samples, the sampling fraction can be computed at each stage, assuming sampling is with equal probability within the stage. A two-stage sample could include selection of n_a primary sampling units from a universe of N_a, and within the ath primary sampling unit, selecting n_{ba} out of N_{ba} units (e.g., households or businesses). In this case, $f_a = (n_a/N_a)$ and $f_{ab} = (n_{ba}/N_{ba})$ and $f_b^* = (\bar{n}_{ba}/\bar{N}_{ba})$ where $\bar{n}_{ba}$ and $\bar{N}_{ba}$ are the mean values of n_{ba} and N_{ba}; the overall sampling fraction would then be $f = f_a(f_b^*)$.

However, many if not most multi-stage samples use selection with probability proportional to size, which makes computing sampling fractions at each stage problematic.

John Hall

See also Finite Population Correction (fpc) Factor; Primary Sampling Unit (PSU); Sampling Interval; Sampling Without Replacement; Strata; Universe

Further Readings

Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample survey methods and theory*. New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

SAMPLING FRAME

A survey may be a census of the universe (the study population) or may be conducted with a sample that represents the universe. Either a census or a sample survey requires a sampling frame. For a census, the frame will consist of a list of all the known units in the universe, and each unit will need to be surveyed. For a sample survey, the frame represents a list of the target population from which the sample is selected. Ideally it should contain all elements in the population, but oftentimes these frames do not.

The quality of the sample and, to an extent, of the survey itself depends on the quality of the sampling frame. Selecting a sampling frame that is of high quality and appropriate both to the population being studied and to the data collection method is a key step in planning a survey. In selecting a sample frame, three questions can be asked: (1) Does it include members of the universe being studied? (2) Is it appropriate for the way the data will be collected? and (3) What is the quality of the frame in terms of coverage, completeness, and accuracy?

Types of Sampling Frames

Major categories of sampling frames are area frames for in-person interviews, random-digit dialing (RDD) frames for telephone survey samples, and a variety of lists used for all types of surveys. Few lists that are used as sampling frames were created specifically for that use. Exceptions are commercially available RDD frames.

The type of frame usually varies with the mode of interviewing, although many frames can be used for multiple modes. Some studies employ multiple frames, either because they use multiple modes of data collection, because no single frame has adequate coverage, or to facilitate oversampling of certain groups.

An in-person survey of households (or individuals living in households) may use multiple levels of frames: an area frame to select a sample of areas where interviews are conducted, and within the areas, lists of addresses compiled by field staff or obtained from commercial sources.

Telephone household surveys may employ RDD frames, directory-based frames, or a combination. Telephone surveys of businesses often use frames developed from telephone directories. Telephone surveys can also use as sampling frames lists from many sources, including government agencies, commercial vendors of lists, associations, and societies. Some of these lists are publicly available, and some can be used only when doing studies for the owner of the list. Examples of publicly available lists include lists of public school districts and schools maintained by the National Center for Education Statistics (there are also commercial frames of districts and schools) and lists of physicians maintained by the American Medical Association. Lists whose use is restricted include those of recipients of government assistance and customers of businesses.

Surveys conducted by regular mail or email often use as frames the same lists (mentioned in the previous paragraph) for telephone surveys. Web surveys could also use these lists as means to contact respondents via regular mail and request that they complete a questionnaire online. Another type of frame for Web surveys comprises one or more Web portals (Web sites that provide links to other Web sites).

Quality Issues

Ideally, the sampling frame will list every member of the study population once, and only once, and will include only members of the study population. The term *coverage* refers to the extent to which these criteria are met. In addition, the frame should be complete in terms of having information needed to select the sample and conduct the survey, and the information on the frame should be accurate.

Needless to say, almost no sampling frame is perfect. Examining the quality of a frame using the criteria discussed in this section may lead to looking for an alternative frame or to taking steps to deal with the frame's shortcomings.

Problems in frame coverage include both undercoverage and overcoverage. *Undercoverage* means that some members of the universe are neither on the frame nor represented on it. Some examples of undercoverage are the following:

1. All RDD landline frames exclude households with no telephone service, and those with only cellular phone service.
2. Frames drawn from telephone directories exclude those households (listed in #1 above) plus those with unpublished and recently published numbers.
3. New construction may be excluded from lists of addresses used as sampling frames for surveys conducted by mail or personal visit.
4. Commercial lists of business establishments exclude many new businesses and may underrepresent small ones.

Frames can also suffer from undercoverage introduced by self-selection bias, as in the case of "panels" recruited for Internet research, even if the panels were recruited from a survey that used a probability sample with a good frame.

Overcoverage means that some elements on the frame are not members of the universe. For example, RDD frames contain nonworking and business telephone numbers, as well as household numbers. A frame may have both undercoverage and overcoverage. For example, to select a sample of students enrolled in a school, one might use a list provided by the school or the district; however, the list might include students who had dropped out or transferred and omit students who had enrolled after the list was compiled.

Frame undercoverage can lead to bias in estimates made from survey data. Overcoverage can lead to bias if ineligible units on the frame are not identified. However, the larger problem with overcoverage is usually one of cost, because ineligibles must be identified and screened out. If the ineligibles can be identified before selecting the sample, it is usually better to eliminate them at that time.

An issue related to coverage is that of duplicates on the frame, which can lead to units having unequal chances of selection. It is best to eliminate duplicates before selecting the sample. If this cannot be done, then the presence of duplicates should be determined for those units that are sampled, so the sample can be properly weighted.

In addition to issues of coverage, a sampling frame should have information that is complete and accurate. For a sampling frame to be complete, it must have enough information so that the sampled units can be identified and located. Further, this information should be accurate. Missing or inaccurate information on the frame can affect the survey's response rate and data collection costs.

John Hall

See also Area Frame; Coverage; Coverage Error; Ineligible; List-Assisted Sampling; Overcoverage; Random-Digit Dialing (RDD); Undercoverage; Universe

Further Readings

Kish, L. (1965). *Survey sampling*. New York: Wiley.
Levy, P., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications* (3rd ed.). New York: Wiley.

Sampling Interval

When a probability sample is selected through use of a systematic random sampling design, a random start is chosen from a collection of consecutive integers that will ensure an adequate sample size is obtained. The *length* of the string of consecutive integers is commonly referred to as the *sampling interval*.

If the size of the population or universe is N and n is the size of the sample, then the integer that is at least as large as the number N/n is called the sampling interval (often denoted by k). Used in conjunction with systematic sampling, the sampling interval partitions the universe into n zones, or strata, each consisting of k units. In general, systematic sampling is operationalized by selecting a random start between 1 and the sampling interval. This random start, r, and every subsequent kth integer would then be included in the sample (i.e., r, $r+k$, $r+2k$, etc.), creating k possible cluster samples each containing n population units. The probability of selecting any one population unit and consequently, the probability of selecting any one of the k cluster samples is $1/k$. The sampling interval and its role in the systematic sample selection process are illustrated in Figure 1.

For example, suppose that 100 households are to be selected for interviews within a neighborhood containing 1,000 households (labeled 1, 2, 3, . . . , 1,000 for reference). Then the sampling interval, $k = 1{,}000/100 = 10$, partitions the population of 1,000 households into 100 strata, each having $k = 10$ households. The random start 1 would then refer to the cluster sample of households {1, 11, 21, 31, 41, . . . , 971, 981, 991} under systematic random sampling.

In practice, the population size may not be an even integer multiple of the desired sample size, so the sampling interval will not be an integer. To determine an adequate sampling interval, one of the following adjustments may be useful.

1. Allow the sample size to be either $(n-1)$ or n. The sampling interval, k, is then chosen so that $(n-1)\times k$ is smaller than N and $n\times k$ is larger than N. Choosing a random start between 1 and k will imply a final sample size of either $(n-1)$ or n units. For example, if a sample of 15 houses is desired from a block containing 100, then $N/n = 100/15 = 6.67$. Choosing a sampling interval of $k = 7$ and allowing the sample size to be either 14 or 15 would then satisfy the requirement: $(15-1)\times 7 = 98 \leq 100$ and $15\times 7 = 105 \geq 100$. In this case, the sampling interval would be $k = 7$; random starts 1 and 2 would yield samples of size 15 while random starts 3 through 7 would yield samples of size 14.

2. Allow circular references in the selection. In this case, the sampling interval is conveniently defined to be any integer no larger than N. A random start from 1, 2, . . . , N is chosen and that unit along with every successive kth unit is selected—if the numbers being selected surpass N, simply continue counting from the beginning of the list as though the population identifications are arranged in a circular fashion. Continue selection until the desired sample size is reached. For example, suppose a sample of 5 households is to be selected from a block having 16 households; a sampling interval of 3 and a random start of 2 results in sampling households 2, 5, 8, 11, 14, 1 (identification number 17 exceeds the population size by 1, so the first element in the list is selected).

3. Use fractional intervals. This approach combines the last approach with a modified computation of sampling interval. For example, suppose there were 200 high school students within a particular graduating

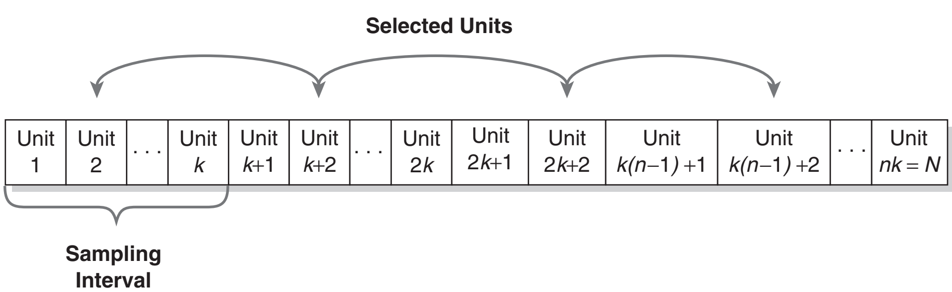

Figure 1 Illustration of sampling interval and its implication for systematic random sampling

Population units are ordered according to identification numbers: 1, 2, . . . $nk = N$. The size of the sampling interval is k units from which a random start is selected. In this case, '2' is selected (circled) for the sample implying that every subsequent kth unit (i.e., 2, $k+2$, $2k+2$, . . .), as shown by upper arrows, is selected for the sample.

class of which a sample of 16 was desired. The corresponding (fractional) sampling interval is $k = 200/16 = 12.5$.

Trent D. Buskirk

See also Cluster Sample; n; N; Random Start; Systematic Sampling; Weighting

Further Readings

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.
Kish, L. (1995). *Survey sampling*. New York: Wiley.
Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Sampling Pool

Sampling pool is a survey operations term, one that statisticians sometimes refer to as the *designated sample size*, which was proposed by Paul J. Lavrakas in the 1980s to refer to the set of elements selected from a sampling frame that may or may not all be used in completing data collection for a given survey project. The value of using this term is to be able to have a unique term to differentiate the sampling pool that a researcher *starts with* from the final sample (i.e., the *final sample size*) the researcher *finishes with*. Traditionally, survey researchers have used the word *sample* to refer to both the final number of completed interviewers a survey is striving to attain and the number of elements used to gain those completed interviews. Because noncontacts, nonresponse, and other reasons (e.g., ineligibility) cause many elements in a sampling pool to not end as completed interviews, the final sample is essentially always smaller in size than the sampling pool and, in many cases, is substantially smaller, for example, 1/10 or 1/20 the size of the sampling pool.

For example, if researchers have estimated that they will need 10,000 telephone numbers for a random-digit dialing (RDD) survey that has a goal of completing 800 interviews, the survey call center that does the interviewing may not need to activate all of those numbers during the data collection period. That is, their processing of the RDD numbers toward the goal of 800 completions may be more efficient than expected, and they may not need to activate all the numbers that were selected for the sampling pool. To allow the sample coordinator the ability to closely manage the sampling, typically all the numbers in the sampling pool will be divided into *sample replicates*. If, for example, the sampling pool contained 10,000 RDD numbers made up of 100 replicates, then each replicate would contain a random subset of 100 of the 10,000 numbers. The sample coordinator may start data collection by releasing half of the replicates (thus a random half of the numbers in the sampling pool) on the first day of the survey's field period. Then the coordinator might observe for the next day or two how efficiently the interviewers are able to process the released numbers in achieving completed interviews. If the efficiency is better than the researchers anticipated, then the coordinator may only need to release another 30 replicates (another 3,000 numbers) to attain the final sample size goal of completed interviews for this survey. Thus, in this example, 2,000 numbers (i.e., 20 replicates) from the sampling pool would never be dialed by interviewers. Of further note, these unreleased numbers would not be considered in any response rate calculations the researchers performed after data collection was completed.

To estimate the size of the sampling pool a given telephone survey needs, Lavrakas advised use of the following formula:

$$\begin{aligned}&\textit{Estimated Size of the Sampling Pool}\\&\quad = (FSS)/((HR)(1 - REC)(1 - LE))\end{aligned}$$

Here *FSS* stands for the final number of completed interviews the survey must attain; *HR* stands for the hit rate, or the estimated proportion of the sampling pool that will reach residences; *REC* stands for respondent exclusion criteria, or the estimated proportion of households that will be deemed ineligible for the particular survey; and *LE* stands for loss of eligibles, or the estimated proportion of eligibles that will end as nonresponders. For example, if in an RDD survey 1,000 completed interviews are desired and the HR is known to be about .65 (65% of the numbers will be households), REC is .05 (5% of households will not have an eligible adult in residence), and LE is .75 (75% of the eligible household will not complete an interview due primarily either to refusals or noncontacts), the estimated size of the sampling pool needed to complete this survey would be (1000)/((.65)(1 − .05)(1 − .75)) or 6,478 RDD numbers. Thus to be on the safe side, the researchers might decide to start with a sampling pool of 8,000 numbers.

Of note, although Lavrakas proposed the use of this term in reference to telephone surveys, the term can be applied to any survey sampling mode—in-person, mail, or Internet.

Paul J. Lavrakas

See also Elements; Nonresponse; Response Rates; Sample; Sample Replicates; Sample Management; Sample Size; Sampling Frame

Further Readings

Lavrakas, P. J. (1987). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.

Sampling Variance

Sampling variance is the variance of the sampling distribution for a random variable. It measures the spread or variability of the sample estimate about its expected value in hypothetical repetitions of the sample. Sampling variance is one of the two components of sampling error associated with any sample survey that does not cover the entire population of interest. The other component of sampling error is coverage bias due to systematic nonobservation. The totality of sampling errors in all possible samples of the same size generates the sampling distribution for a given variable. Sampling variance arises because only a sample rather than the entire population is observed. The particular sample selected is one of a large number of possible samples of the same size that could have been selected using the same sample design. To the extent that different samples lead to different estimates for the population statistic of interest, the sample estimates derived from the different samples will differ from each other.

The positive square root of the sampling variance is called the *standard error*. For example, the square root of the variance of the sample mean is known as the *standard error of the mean*. The sample estimate and its standard error can be used to make inferences about the underlying population, for example, through constructing confidence intervals and conducting hypothesis testing. It is important to note, however, that sampling variance is measured about the expected value of the statistic under the sample design rather than the true population value. Therefore, inferences based on sampling variance do not reflect sampling biases or any possible nonsampling errors.

Under probability sampling, the sampling variance can be estimated using data collected from the sample. The estimation methodology for the sampling variance should take into account both the sample design and the estimation method. For standard sampling designs and estimators, standard variance estimation formulae are available. In his book *Introduction to Variance Estimation*, Kirk M. Wolter discussed nine basic sampling designs and their associated variance estimators.

In many samples, however, data are collected from individuals or organizations using complex sample designs that typically involve unequal selection probabilities, sample stratification, clustering, and multistage sampling. For such complex sample designs, although it is possible to produce unbiased point estimates by using proper sample weights, it is generally not possible to estimate appropriate sampling variances using standard estimation methods. In fact, for many complex sample designs and estimators, exact algebraic expressions of the sampling variances are not available, and hence there are no direct analytic methods for producing unbiased variance estimates.

One general approach to approximating the sampling variance of an estimator is to use sample replication methods such as jackknife and balanced repeated replication. An alternative approach is to approximate the estimator analytically using Taylor series expansion and then compute the variance of the linearized estimator. Statistical software packages that specialize in complex variance estimation include SUDAAN, WesVar, and STATA, among others.

Y. Michael Yang

See also Balanced Repeated Replication (BRR); Jackknife Variance Estimation; Nonsampling Error; Probability Sample; Sample Size; Sampling Bias; Sampling Error; Standard Error; Stata; SUDAAN; Taylor Series Linearization; Variance Estimation; WesVar

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer.

Sampling Without Replacement

In sampling without replacement, each sample unit of the population has only one chance to be selected in the sample. For example, if one draws a simple random sample such that no unit occurs more than one time in the sample, the sample is drawn *without replacement*. If a unit can occur one or more times in the sample, then the sample is drawn *with replacement*. The same concept applies to other types of sample designs. For example, in multi-stage sampling the first-stage sampling units (primary sampling units) can be drawn from strata without replacement or with replacement.

In the case of simple random sampling, the estimator of the population mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ where n is the sample size and y_i is the value of a study variable for the ith unit in the sample. The estimator of the variance of the sample mean is

$$s_{\bar{y}}^2 = \frac{(1-f)}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{(n-1)}.$$

The term $(1-\mathrm{f})$ equals $(1-\frac{n}{N})$ where N is the population size and is known as the *finite population correction* when sampling without replacement. Thus, if the sample size is large relative to the population size (e.g., 10% or higher), the finite population reduces the variance of the sample mean when compared to a simple random sample of the same size, n, drawn with replacement. The remainder of this entry provides a more detailed background for sampling without replacement.

A sample without replacement is usually defined as a subset of a finite population. A finite population of elements is represented by $U = \{1, \ldots, N\}$. A *sample* is a vector $s = (s_1, \ldots, s_N)^T$; T is used to denote matrix transposition, where $s_k \in \mathrm{N}$ (N is the set of natural numbers). If $s_k \in \{0, 1\}$ for all $k \in U$, then the sample s is without replacement. In other words, a sample without replacement is a vector of indicator variables. Furthermore, in a sample without replacement

$$s_k = \begin{cases} 1 & \text{if the element } k \text{ is in the sample,} \\ 0 & \text{if the element } k \text{ is not in the sample,} \end{cases}$$

for all $k \in U$. On the other hand, if s_k is any natural number, then s is a sample with replacement. The sample size of s is defined by

$$n(s) = \sum_{k\in U} s_k.$$

Let D be a set of samples. Moreover, let P be a probability mass function with support D. This probability mass function P is a *sampling design*. That means that P is a function from support D to (0,1) such that $P(s) > 0$ for all $s \in D$ and $\sum_{s\in D} P(s) = 1$. If $D \subset \{0, 1\}^N$, then P is a *sampling design without replacement*. The term $D \subset \{0,1\}^N$ means that any $s \in D$ is a vector of 0s and 1s; it is a sample without replacement. Furthermore, if $n(s) = n$, it is constant for all $s \in D \subset \{0,1\}^N$, then P is a sampling design without replacement and with fixed sample size.

For clarifying concepts, the following example is provided. Let $U = \{1, 2, 3\}$ be a population of size three. Thus,

$$\{0,1\}^3 = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}$$

is the set of all possible samples without replacement. For example, the vector (1,0,1) denotes that the first and third elements of the population U are in the sample. Moreover, $D_2 = \{(0,1,1), (1,0,1), (1,1,0)\}$ is the set of samples without replacement and with fixed sample size equal to 2. A possible sampling design P with support D_2 is

s	$(0,1,1)$	$(1,0,1)$	$(1,1,0)$
$P(s)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{4}$

Now, let P be a sampling design with support D and let $S = (S_1, \ldots, S_N)^T$, $S_k \in N$ for all $k \in U$, be a random vector such that $\Pr(S = s) = P(s)$ for all $s \in D$. Sometimes, D is also referred to as the support of this random vector S; it is the set of possible values of S. Consequently, the expectation of the random vector S is $\pi = E(S) = \sum_{s\in D} P(s)s$. Observe that π is a vector of the same dimension of S. Furthermore, if the sampling design P is a sampling design without replacement, then π is the vector of the first inclusion probabilities. Observe that, in the case of sampling without replacement, the kth component S_k of S only takes values in {0,1}. Hereby, the kth component of the vector of expected values π can be expressed as $\pi_k = E(S_k) = \sum_{s\in D} P(s)s_k = \sum_{\{s\in D: s_k=1\}} P(s)$, the known definition of the first inclusion probability of the sample unit $k \in U$. Now, the matrix of second moments of

S is $\Pi = E(SS^T) = \sum_{s \in D} P(s)ss^T$. Once more, if the sampling design P is without replacement, then Π is the matrix of second inclusion probabilities, with $\pi_{kk} = \pi_k$ for all sample unit $k \in U$.

The best-known sampling without replacement designs are the Bernoulli sampling, simple random sampling without replacement, and Poisson sampling. In the Bernoulli sampling, the support is $D = \{0,1\}^N$; the set of the all possible subsets of U. Moreover, its sampling design is $P(s) = \theta^{n(s)}(1-\theta)^{N-n(s)}$, where the $S_1, \ldots, S_N$ are independent and identically distributed Bernoulli random variables with parameter $\theta \in (0, 1)$. In the case of Poisson sampling, the support also is $D = \{0, 1\}^N$ and its sampling design can be expressed by $P(s) = \prod_{k \in U} \pi_k^{s_k}(1-\pi_k)^{1-s_k}$ for all $s \in D$, where the $S_1, \ldots, S_N$ are independent, distributed Bernoulli random variables with parameters $\pi_1, \ldots, \pi_N \in (0, 1)$ respectively.

Bernoulli sampling and Poisson sampling are designs that produce variable size samples. If the support D is restricted to samples of fixed size, say n, then the Bernoulli sampling turns into the simple random sampling without replacement. In this case the support is $D_n \subset \{0,1\}^N$; the set of all the possible subsets of U but with fixed sample size n. The sampling design for simple random sampling without replacement is

$$P(s) = \frac{1}{\binom{N}{n}}.$$

However, the $S_1, \ldots, S_N$ are not independent; the correlation between S_i and S_j is

$$-\frac{1}{N-1}, \ i \neq j \in U.$$

On the other hand, if Poisson sampling also is restricted to a support with samples of fixed size, then the conditional Poisson sampling is obtained.

José Elías Rodríguez

See also Finite Population Correction (fpc) Factor; Probability Proportional to Size Sampling (PPS); Sample; Sample Design; Sampling With Replacement; Simple Random Sample

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Tillé, Y. (2006). *Sampling algorithms*. New York: Springer.

SAS

SAS (pronounced "sass") is the name of one of the world's largest software development corporations. Originally an acronym for "statistical analysis software," SAS was created by Jim Goodnight, John Sall, and other researchers at North Carolina State University in the early 1970s. What began as a locally developed set of programs for agricultural research quickly became so popular that in 1976 the SAS Institute was formed in Raleigh, North Carolina, to meet the growing demand. The company immediately formed an alliance with IBM and created the first SAS Users Group International (SUGI), which continues to provide assistance to SAS users, distribute newsletters, maintain a popular Web site and hold conferences throughout the world.

Within 5 years, SAS outgrew its original site and moved to its current campus in Cary, North Carolina. By this time, it was installed in thousands of sites around the world. Within 10 years, not only was SAS installed on 65% of all mainframe sites, but partnerships had been established with Microsoft and Apple as the personal computer revolution began. Throughout the 1990s and early 2000s, SAS has been the recipient of many prestigious awards for its technical accomplishments. Annual revenues in 1976 were $138,000; by 2006 they were $1.9 billion.

The features of SAS software cover a large family of products with applications in the government, academic, and private sectors. The characteristics most used by survey research professionals are common to most data analysis software, although the implementation can be very different from one software package to another. First, SAS has the ability to read electronically stored data in almost any format from almost any medium. Second, it has an enormous array of data transformation options with which to recode existing variables and create new ones. Third, SAS has an unlimited number of data analysis procedures from commonly used procedures to the most exotic analysis and the capability of creating user-developed applications that can be implemented from within existing SAS applications. Fourth, output can be in tabular or graphical form. Finally, SAS has a full-featured macro language with which to control data processing and output.

James Wolf

Further Readings

SAS: http://www.sas.com

Satisficing

The notion of satisficing is consistent with cognitive theory articulated by Roger Tourangeau, Lance Rips, and Kenneth Rasinski that survey respondents must execute four stages of cognitive processing to answer survey questions optimally. Respondents must (1) interpret the intended meaning of the question, (2) retrieve relevant information from memory, (3) integrate the information into a summary judgment, and (4) map the judgment onto the response options offered. When respondents diligently perform each of these four steps, they are said to be *optimizing*. However, instead of seeking to optimize, respondents may choose to perform one or more of the steps in a cursory fashion, or they may skip one or more steps altogether. Borrowing Herbert Simon's terminology, Jon Krosnick has referred to this behavior as *satisficing* in his seminal paper published in 1991.

Whereas some people may begin answering a questionnaire without ever intending to devote the effort needed to optimize, others might begin to answer a questionnaire with the intention to optimize, but their enthusiasm may fade when they face a long questionnaire or questions that are difficult to understand or answer. As they proceed through the questionnaire, these respondents may become increasingly fatigued, distracted, and uninterested. But even after motivation begins to fade, the fatigued and unmotivated respondent is nevertheless expected to continue to provide answers to questions with the implicit expectation that he or she will answer each one carefully. At this point, a respondent may continue to expend the effort necessary to provide optimal responses or may choose instead to answer questions more superficially, expending less mental energy and short-cutting the steps necessary for optimal answering; in other words, they might satisfice.

Forms of Satisficing

Respondents who devote less-than-optimal effort to the task of answering questions can engage in weak or strong satisficing. Weak satisficing occurs when a respondent performs all four cognitive steps but performs one or more of these less carefully or attentively than is needed to optimize. A respondent implementing weak satisficing may be less thoughtful in inferring the intended meaning of a question, less thorough in searching memory for all relevant information, less balanced in integrating the retrieved information into a summary judgment, and more haphazard in selecting the appropriate response option from the list offered. Strong satisficing occurs when a respondent skips the retrieval and judgment steps altogether and seeks merely to identify a plausible answer based on cues provided by the question, without reference to any internal psychological cues directly relevant to the attitude, belief, or event of interest to the researcher. If no cues pointing to such an answer are immediately evident in a question, a satisficing respondent may choose a response at random. Strong satisficing allows a respondent to provide a reasonable and seemingly defensible answer while applying very little effort. Rather than making a sharp distinction between weak and strong satisficing, Krosnick proposes that an individual's response to any given question can fall somewhere along a continuum ranging from optimizing at one end to strong satisficing at the other.

Conditions Under Which Satisficing Is Likely

Krosnick has hypothesized that the likelihood a survey respondent will satisfice is a function of the respondent's ability to perform the cognitive tasks of optimizing, the respondent's motivation to perform the tasks, and the difficulty of the tasks. Satisficing should be more common when the respondent has less ability to optimize, when the respondent is less motivated to optimize, and when the tasks are more difficult.

Ability

A key aspect of ability is the respondents' level of cognitive skills to perform the complex mental operations required by optimizing. Satisficing theory defines cognitive skills as the ensemble of abilities needed to interpret questions, retrieve information from memory, integrate that information into a summary judgment, and express it verbally. People with limited skills at language interpretation, knowledge retrieval, retention and manipulation of information in working memory, judgment, and verbal expression are presumably least able to optimize and are therefore especially likely to satisfice instead. In contrast, people with strong skills in these areas should find it easy to execute the steps of optimizing and may therefore be especially likely to do so. Thus, differences between respondents in levels of cognitive skills should differentiate satisficers from optimizers.

Motivation

There are many potential sources of motivation to optimize when answering questionnaires, including the respondent's need for cognition, the extent to which the question topic is personally important to the respondent, the degree to which the respondent is held accountable for the answers he or she provides, the number of previous questions he or she has answered in a questionnaire, and more. More motivation presumably enhances the likelihood of optimizing.

Task Difficulty

Task difficulty is a feature of a question and depends on how much mental work is required to accomplish the task set out by the question. For example, interpreting the meaning of a question can be especially challenging if the words in it have multiple meanings, so respondents are forced to use linguistic context to infer the intended meanings of the words. Likewise, extraneous events occurring during questionnaire completion may distract a respondent from thinking about a question, making the task more difficult. These and other sources of difficulty may decrease the likelihood that respondents will perform all four steps of optimizing fully.

In his original formulation, Krosnick raised the possibility that ability, motivation, and task difficulty may interact to regulate satisficing. The interaction could manifest in one of two ways: (1) Optimizing might be the default approach that respondents take to answering questionnaires, so satisficing may be more likely when low levels of ability co-occur with low level of motivation, high task difficulty, or both, or (2) satisficing might be the default approach that respondents take to answering questionnaires, so optimizing might occur only when high levels of ability co-occur with high motivation, low task difficulty, or both.

Satisficing has been posited to at least partly explain several response effects, including acquiescence effects, response order effects, no opinion option effects, and nondifferentiation in answering batteries of rating scale items. Respondents inclined to satisfice may employ a number of different strategies to select seemingly legitimate answers while expending minimal cognitive effort. If offered a closed-ended question with categorical response options (e.g., *Which of the following experiences frustrates you most often: waiting in long lines, accidentally dropping things, or forgetting things you need to remember?*), a respondent may choose the first response option that seems reasonable, rather than reading the entire list of choices and thinking carefully about each option individually; this situation yields response order effects. When questions ask whether a set of statements are true or false or whether respondents agree or disagree with specific statements, a confirmatory bias in retrieval and reasoning would lead satisficing respondents to agree with assertions rather than disagreeing with them (acquiescence bias). And when offered a question with an explicit "don't know" response option, respondents might pick it to avoid answering the question substantively. All of these strategies allow respondents to appear to answer questions legitimately without having to think about their topics at all.

A good amount of evidence has accumulated that is consistent with satisficing theory's contention that response order, acquiescence, and Don't Know response option effects are more common under the three conditions described earlier. For example, these effects tend to be stronger among respondents with limited cognitive skills, when questions are more difficult to comprehend, among respondents who are not motivated to think, when questions are placed later in a long questionnaire, and under many more such conditions.

Questionnaire design features that minimize cognitive burden for the respondent (e.g., using commonly used words and avoiding jargon, keeping the questionnaire to a reasonable length, and more) can help to increase the likelihood of optimal responding to survey questions.

Sowmya Anand

See also Acquiescence Response Bias; Cognitive Aspects of Survey Methodology (CASM); Nondifferentiation; Questionnaire Length; Respondent Burden; Respondent Fatigue; Response Bias; Response Order Effects

Further Readings

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Screening

Screening is the process by which elements sampled from a sampling frame are evaluated to determine whether they are eligible for a survey. Ideally, all members of the sampling frame would be eligible, but eligibility information is often not available prior to constructing the frame. In this case, the sampling frame must be narrowed to include only eligible sample members by subsectioning the frame, matching it against an external administrative data source, or collecting eligibility information directly from a sampled respondent or a proxy for that respondent.

Screening Types

When the sample frame is subsectioned or matched against an external administrative data source, this is referred to as *passive screening* because a respondent is not directly involved in the process. Passive screening uses existing data to determine who, from a sampling frame of individuals, establishments, or other, is likely eligible for a survey. For instance, a survey of pediatric specialty hospitals in the western United States may begin with a list of all hospitals across the United States. Based on the original list itself, or another that has been merged with the original, the list can be narrowed down to those hospitals located in western states. The list could be further screened to include only those with a pre-designated classification of being a pediatric hospital.

When eligibility information is obtained directly from a respondent or proxy, this is referred to as *active screening*. Active screening involves direct contact with potentially eligible respondents and is typically undertaken when the eligibility criteria are not available from the sample frame. In this scenario, potentially eligible respondents are contacted in person, by phone, by Web, or by mail to determine their eligibility through a short screening interview (or "screener"). A variety of eligibility criteria may be evaluated by actively screening respondents. Considerations may include age, race, education, income, or geographic location, among others, depending on the purpose of the survey. Active screening is also often done to identify rare or difficult-to-locate populations. For example, a household survey of Hispanic single mothers may include an active screening component to determine the age, ethnicity, and relationships among all household members to make a determination as to whether anyone in that household is eligible as "Hispanic single mother." Households would be contacted and asked questions related to these demographic characteristics, and only those meeting the eligibility criteria would be retained for possible participation in the survey.

Active screening can be completed using several different modes and at a time different from data collection for the main survey. Often, but not always, the screening takes place in the same mode as the main survey interview. For instance, in a general population household survey, an interviewer may visit the household in person to administer the screener to the person answering the door. The screener may be designed to identify household members with a certain characteristic or within a certain age range. From the screener results, the main interview respondent or respondents can be selected as specified by the sampling criteria from the pool of eligible household members, and the interview can then be conducted in person with the selected respondent or respondents. Screening can also take place over the phone, Web, or by mail, which are often more cost-effective techniques compared to in-person screening, though each has associated sources of potential error.

In random digit dialing surveys, active versus passive screening for geographic location is becoming increasingly necessary. The portability of telephone numbers and the growth rate of cell phone only households who can take their cell phone number with them when they move have made it more difficult to determine with which geographic area a phone number is associated. Active screening for geographic eligibility is often difficult because there is some degree of error associated with respondents' ability and willingness to report their location, even when detailed descriptions of the study geography are provided.

Other sources of error are associated with both passive and active screening. In passive screening, false positive eligibles may be retained in the frame if the data related to eligibility indicate eligibility when the case is truly ineligible. Conversely, false negatives might be erroneously excluded if the sample frame incorrectly suggests the case is ineligible or eligibility information is missing. Active tracing is subject to error from false positives and false negatives as well. If a respondent advertently or inadvertently provides incorrect eligibility information about himself or herself, the family, or the establishment, the pool of eligible respondents may be incorrect. To the extent this error is correlated with some attribute important to the study purpose, bias may be introduced that could negatively affect the quality of the final survey estimates.

Screening Techniques

Because screening is often conducted very early in the interview process, it is vital that screening techniques are designed to foster, rather than discourage, participation. Screening questions and interviews should be as brief and to the point as possible. The screener language should not be written in a way that might bias the response decision process. For instance, a survey on charitable donations among wealthy people may include a screener that simply obtains the respondent's basic demographic information and general income range rather than collecting a great deal of information about the respondent's sources of wealth and charitable behaviors. Such detailed questions can be administered in the main survey once the correct respondents have been identified and the specifics of the survey topic and benefits have been explained. A brief and persuasive screening approach can help limit the potential for nonresponse bias by maximizing the rate of successful screening. However, it is paramount that those being screened remain unaware of which responses will screen them "in" or "out." Were this not to happen, some respondents purposely would provide incorrect answers in order to lead to a screening result they preferred, which likely would bias the screening process.

In many cases, it may be difficult even to conduct screenings. Respondents may be unavailable or unwilling to complete a screener. When eligibility for these screening nonrespondents is not obtained, they are referred to as being of "unknown eligibility." When computing outcome rates, there are many methods for treating cases of unknown eligibility, but it is always the goal of a screening to end the survey's field period with the number in this group minimized.

Screening and Response Rates

In multi-stage sample designs, the ability to successfully screen potential households or respondents is calculated as a *screening rate*. The screening rate is often multiplied by the main interview response rate to obtain an overall response rate for a survey. Multiplying the rates assumes that the distribution of eligible persons in nonrespondent sample households is the same as in the respondent sample households. The American Association for Public Opinion Research recommends that some investigation of this assumption be conducted if this overall response rate computation is utilized.

Joe Murphy

See also Elements; Eligibility; Geographic Screening; List-Assisted Sampling; Multi-Stage Sample; Number Portability; Proxy Respondent; Rare Populations; Response Rates; Sampling Frame; Unknown Eligibility

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author. Retrieved April 17, 2008, from http://www.aapor.org/uploads/standarddefs_4.pdf

Ezzati-Rice, T., Coronado, V., Frankel, M., Hoaglin, D., Loft, J., & Wright, R. (1999). *Estimating response rates in random-digit-dialing surveys that screen for eligible subpopulations*. Paper presented at the International Conference on Survey Nonresponse, Portland, OR.

Retrieved December 21, 2006, from http://www.jpsm.umd.edu/icsn/papers/EzzatiRiceCoronado.htm

Massey, J. T. (1995). Estimating the response rate in a telephone survey with screening. *Proceedings of the Section on Survey Research Methods* (Vol. 2, pp. 673–676). Alexandria, VA: American Statistical Association.

Meegama, N., & Blair, J. (1999). The effects of telephone introductions on cooperation: An experimental comparison. *Proceedings of the Section on Survey Research Methods* (pp. 1029–1031). Alexandria, VA: American Statistical Association.

Smith, T. W. (2003). *A review of methods to estimate the status of cases with unknown eligibility*. Retrieved April 17, 2008, from http://www.aapor.org/uploads/erate.pdf

Snyder, D. C., Sloane, R., Lobach, D., Lipkus, I., Clipp, E., Kraus, W., et al. (2004, October). Agreement between a brief mailed screener and an in-depth telephone survey: Observations from the Fresh Start Study. *Journal of the American Dietetic Association*, *104*, 1593–1596.

Seam Effect

The seam effect, also called the *seam bias*, a phenomenon specific to longitudinal panel surveys, refers to the tendency for estimates of change, as measured across the "seam" between two successive survey administrations (or "waves"), to far exceed change estimates that are measured within a single survey wave—often by a factor of 10 or more. Seam effects have been found in virtually every panel survey examined, regardless of the characteristics under study, the data collection methods, or the length of the recall period. Seam bias almost always signals the presence of serious measurement error, which can severely compromise the statistical utility of estimates of change. A considerable amount of research over the past two decades has documented the existence of seam effects in longitudinal surveys and also has shed light on their essential nature—too little change is observed within the reference period of a single interview wave, and too much is observed at the seam.

Figure 1 presents a typical seam bias profile. It shows month-to-month transitions in reported receipt of Food Stamps and Social Security retirement benefits from the first three interview waves of the 1984 panel of the U.S. Census Bureau's Survey of Income and Program Participation (SIPP). SIPP waves occur at 4-month intervals and collect data about the preceding 4-month period; thus Months 4 and 5, and

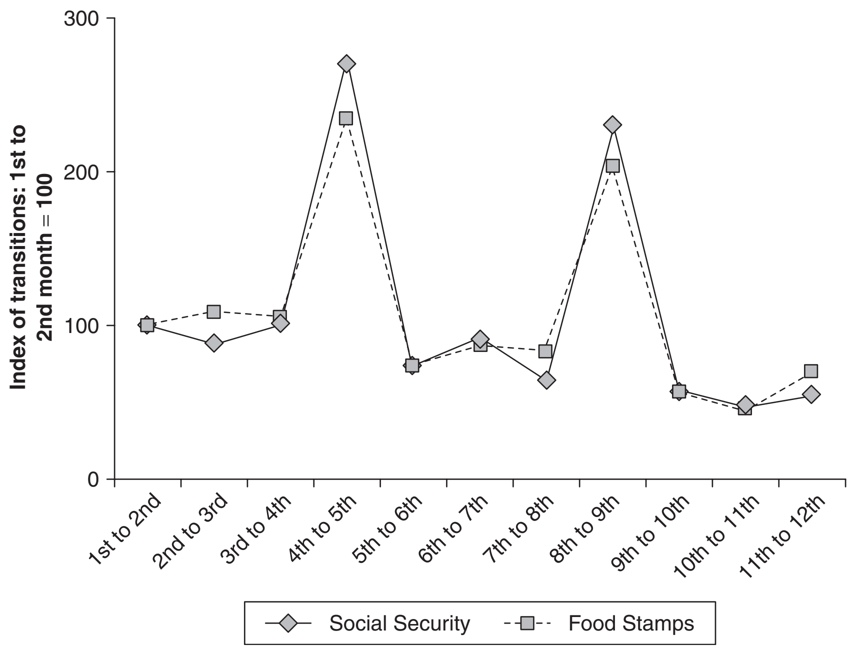

Figure 1 Survey of income and program participation month-to-month transition rates for receipt of social security benefits and food stamps

Source: Adapted from Burkhead and Coder, 1985, pp. 355–356.

Months 8 and 9, comprise the "seams" between Waves 1 and 2 and Waves 2 and 3, respectively, which are reflected by the large spikes in Figure 1.

Many factors have been cited as potential contributors to seam effect phenomena, including the following:

- *Data processing actions*—for example, strategies for assigning missing values and errors in linking cases across interview waves can create spurious transitions at the seam.
- *Interviewer, coder, or respondent inconsistencies*—any kind of interviewer error or inconsistency across successive survey waves is a possible cause of seam bias, as are coder inconsistencies in classifying open-ended questions and respondent inconsistencies in applying labels to phenomena of interest.
- *Self or proxy response status*—spurious change at the seam may result from the fact that respondents can change across successive waves of panel surveys; questionnaire design—unlike most response errors, seam effects characterize phenomena (i.e., month-to-month changes) that generally are not measured directly from respondents' reports but rather are derived in the analysis stage from those data.
- *Memory issues*—memories for more recent portions of the response period of one wave are likely to be of different quality and to result from different recall strategies (e.g., direct recall vs. estimation), as compared to memories for the most remote portion of the response period of the subsequent wave.
- *Satisficing*—in response to a difficult or burdensome recall task, respondents may adopt short-cut strategies such as constant wave responding, in which the same answer is reported for all months of an interview wave's reporting period.

Most evidence, however, discounts the relative importance of the initial, more "extrinsic" factors in the preceding list and suggests instead that questionnaire design, respondent memory issues, and recall strategies play the predominant roles in producing seam effects.

One approach to the amelioration of seam effects is through statistical adjustment after the data have been collected. However, different data collection methods have also been shown to produce different seam effects. In particular, use of *dependent interviewing* has been shown to substantially reduce, although not eliminate, seam bias.

Mario Callegaro and Jeffrey C. Moore

See also Aided Recall; Coding; Dependent Interviewing; Interviewer-Related Error; Longitudinal Studies; Measurement Error; Panel; Questionnaire Design; Record Check; Respondent-Related Error; Satisficing; Unaided Recall; Wave

Further Readings

Burkhead, D., & Coder, J. (1985). Gross changes in income recipiency from the Survey of Income and Program Participation. In *Proceedings of the Social Statistics Section* (pp. 351–356). Washington, DC: American Statistical Association.

Callegaro, M. (2008). Changes in seam effects magnitude due to changes in question wording and data collection strategies. An analysis of labor force transitions in the PSID. In American Statistical Association (Ed.), *Proceedings of the 62nd annual conference of the American Association for Public Opinion Research*, Anaheim, CA.

Moore, J. C., Bates, N., Pascale, J., & Okon, A. (2008). Tackling seam bias through questionnaire design. In P. Lynn (Ed.), *Methodology of longitudinal surveys*. Hoboken, NJ: Wiley.

SEGMENTS

Segments is a term for sample units in area probability sampling (a specific kind of cluster sampling). Most often, segments are sample units in the second stage of area probability sampling and are more formally referred to as *secondary sampling units* (SSUs).

As second stage sample units, segments are neighborhoods or blocks (either census-defined or practically defined by field workers) within the selected *primary sampling units*, which are often counties or whole metropolitan areas). Occasionally, segments can refer to the first-stage sample units or to larger areas than neighborhoods or blocks, such as entire census tracts or even counties, but this entry describes them in their more common usage as second-stage sample units.

Individual units (often housing units, or sometimes clusters of housing units) within the selected segments are selected for inclusion in the sample. Traditionally, field workers are sent out to the selected segments to list every housing unit. New lists are becoming available for segments in urban areas built from postal address lists. These lists are not yet available in rural areas, but as rural addresses in the United States get converted to city-style addresses for 911 reasons, the postal address lists available commercially will continue to increase their coverage.

Segments are usually defined only within primary sampling units that have been selected in the first stage of sample selection. Segments are designed to be as contiguous as possible because this reduces interviewer travel between selected units, but if selected using census data, consecutive-numbered blocks may not be strictly contiguous. There are two key decisions to be made in defining segments. The first is how large to make the segments, and the second is how many segments to select within each primary sampling unit.

Deciding how large to make the segments involves a trade-off of survey cost versus variance. Under traditional listing, larger segments will cost more to list every housing unit. Larger segments will also necessitate more travel between individually selected sample units, which increases survey costs. However, smaller segments result in a more homogenous sample, as measured by larger intraclass correlations, often represented by the Greek letter *rho* (ρ). Larger rho values reduce the *effective sample size*, which results in more variance. The rho differs for each variable, depending on how similar people who live near each other are on any particular characteristic. As a general rule, socioeconomic characteristics (e.g., income) tend to have higher rho values than do behavioral variables. This is because people who live near each other tend to have similar financial situations, but even people in similar financial situations tend to have different opinions and behaviors.

Morris H. Hansen, William N. Hurwitz, and William G. Madow suggested a logarithmic relationship between the average cluster size and rho:

$$\rho = a(\bar{N})^m,$$

where a and m are different parameters for different variables, and $\bar{N}$ represents the average segment size. The assumption in the formula is that $m < 0$ so that as the average cluster size increases, the value of rho will decrease. This is explained by the fact that smaller areas tend to be more similar than larger areas (e.g., one neighborhood vs. an entire city).

Deciding how many segments to select within each primary sampling unit also involves a trade-off of survey cost versus variance. Selecting more segments will reduce the average number of selected third-stage units per segment and, therefore, the average number of cases per segment. Having more cases per segment allows the rho value to affect the sample estimates more.

The variability of sample estimates depends on the number of cases per segment and the value of rho. In fact, there is a well-known approximation to the relationship of the rho value given here:

$$n(eff) \approx \frac{n}{[1 + \rho(\bar{b} - 1)]},$$

where $\bar{b} = n/(\#SSUs)$ is the average number of cases per segment. Note that if ρ is 0 or $\bar{b} = 1$ (simple random sampling has no clustering), the effective sample size is the same as the total sample size g. However, the value of rho is generally positive, so the effective sample size is generally less than the total sample size.

Steven Pedlow

See also Area Probability Sample; Case; Clustering; Cluster Sample; Effective Sample Size; Multi-Stage Sample; Primary Sampling Unit (PSU); ρ (Rho); Survey Costs; Simple Random Sample; Unit; Variance

Further Readings

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: Wiley.

Self-Administered Questionnaire

A self-administered questionnaire (SAQ) refers to a questionnaire that has been designed specifically to be completed by a respondent without intervention of the researchers (e.g., an interviewer) collecting the data. An SAQ is usually a stand-alone questionnaire though it can also be used in conjunction with other data collection modalities directed by a trained interviewer. Traditionally the SAQ has been distributed by mail or in person to large groups, but now SAQs are being used extensively for Web surveys. Because the SAQ is completed without ongoing feedback from a trained interviewer, special care must be taken in how the questions are worded as well as how the questionnaire is formatted in order to avoid measurement error.

A major criterion for a well-designed SAQ is proper wording and formatting of the instructions, the questions, and the answer categories. Don Dillman has succinctly stated that the goal of writing a survey question for self-administration is to develop a query that every potential respondent will interpret in the same way, be able to respond to accurately, and be willing to answer. Dillman also describes a variety of conditions that need to be considered when writing questions for an SAQ. Because respondents usually will have no one to ask for clarification, the SAQ must be completely self-explanatory. However, experience shows that respondent creativity knows no bounds for misunderstanding an SAQ. Thus, all types of alternative interpretations must be considered for each question in order to avoid item nonresponse due to confusion or, a more insidious problem, an improper response based on a misunderstanding of the instructions or question. The structure of the questions is also critical to collecting appropriate data. Closed-ended versus open-ended question structures, ordered versus unordered answer categories, and the anticipated respondent characteristics need to be assessed for their utility for each question.

Another critical criterion for high-quality SAQs is appropriate formatting of the questionnaire, including the size of the font, spacing, navigational aids, use of color, and other aspects. Regardless of how well the questions are worded, a poorly formatted SAQ will result in a variety of problems. A questionnaire should progress in a manner that makes sense to the respondents, not the researcher. Following general principles of conversation is recommended. Questions on similar issues are usually grouped together with topics most relevant to the respondent appearing before others. When topics change, the visual layout should reinforce the switch to a new subject. Consistent use of symbols and other graphics (text boxes, navigational arrows, etc.) can be very useful in helping the respondent follow skip patterns and other deviations from sequential administration. Even for computer-assisted interviewing, the layout of each screen can be as important as that of the printed page. Respondents often begin answering questions without reading instructions unless the layout makes reading those sentences seem particularly important.

The most common form of SAQ is the printed questionnaire delivered to the respondent, usually with a postage-paid envelope for returning the completed questionnaire. As the general public has become more familiar with using personal computers, the SAQ is increasingly being administered using computer programs running either on a stand-alone personal computer or on a Web site designed to present the questions and have respondents enter their answers directly into the database. Research has shown that respondents are more likely to report sensitive or illegal behavior when they are allowed to use a SAQ format rather than during a personal interview on the phone or in person. For this reason SAQs are commonly used to supplement face-to-face interviews when researchers are concerned about social desirability issues.

James Wolf

See also Closed-Ended Question; Computer-Assisted Self-Interviewing (CASI); Computerized Self-Administered Questionnaires (CSAQ); Gestalt Psychology; Internet Surveys; Mail Survey; Open-Ended Question; Questionnaire Design; Question Order Effects; Respondent-Related Error; Sensitive Topics; Social Desirability; Total Design Method; Web Survey; Visual Communication

Further Readings

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

Self-Reported Measure

Self-reported measures are measures in which respondents are asked to report directly on their own behaviors, beliefs, attitudes, or intentions. For example, many common measures of attitudes such as Thurstone scales, Likert scales, and semantic differentials are self-report. Similarly, other constructs of interest to survey researchers, such as behavioral intentions, beliefs, and retrospective reports of behaviors, are often measured via self-reports.

Self-reported measures can be contrasted to other types of measures that do not rely on respondents' reports. For example, behavioral measures involve observing respondents' behaviors, sometimes in a constrained or controlled environment. Similarly, physiological measures like galvanic skin response, pupillary response, and subtle movements of facial muscles rely on biological responses rather than self-report. Measures of other variables, such as weight, height, or cholesterol level could also be assessed without self-report by weighing or measuring respondents or by taking specimens like blood or urine samples, as often is done in health surveys. Finally, implicit measures such as the Implicit Association Test (IAT) and Russell H. Fazio's priming paradigm involve tasks that are not under conscious control and that do not make respondents overtly aware that their attitudes are being measured.

Historically, surveys have almost exclusively made use of self-report measures, but technological advances have made other types of measures more plausible. For example, response latencies can be measured in survey questions either by having interviewers record the length of time respondents take to answer a question. They also can be measured in computer-assisted self-administered and Internet surveys. Biological specimens or measurements can be taken in in-person interviews. Implicit measures like the IAT or priming tasks can be implemented in computer-assisted self-interviewing surveys or Internet surveys.

Although self-report measures are widely used, survey researchers using these measures should be aware that their use is based on the assumptions that respondents are able to answer the questions posed to them and that they are willing to do so, and that these assumptions may not be true. For example, people have limited and often imperfect access to many of their own internal mental processes, and they may therefore not be able to give accurate responses to questions about these processes. However, when asked about internal mental processes, respondents may construct a logical response based on their theories about their mental processes, rather than on actual knowledge of these processes. Thus, respondents will answer questions about these processes, but those answers may not be accurate reflections of the processes themselves. For example, respondents in a survey may willingly answer questions about why they voted for one presidential candidate over another, but those answers may not reflect their actual decision processes. Respondents' self-reports may also be inaccurate because they can be influenced by context, which is demonstrated by research exploring the effect of question order on survey responses. Furthermore, respondents' answers to survey questions may also be inaccurate because of limits to memory or errors in memory.

The use of self-report measures also assumes that respondents are willing to answer researchers' questions. Because being viewed favorably by others is more likely to bring rewards and minimize punishments than being viewed unfavorably, people may sometimes be motivated to construct favorable images of themselves for other people (e.g., for interviewers), sometimes via deceit. Such systematic and intentional misrepresentation by respondents when answering questionnaires has been well documented. For example, people are more willing to report socially embarrassing attitudes, beliefs, and behaviors when their reports are anonymous and when respondents believe researchers have other access to information about the truth of their thoughts and actions. Thus, some people sometimes distort their answers to questions in order to present themselves as having more socially desirable attitudes, beliefs, or behavioral histories, and people's reports may therefore be distorted by social desirability bias.

Allyson Holbrook

See also Behavioral Question; Cognitive Aspects of Survey Methodology (CASM); Computer-Assisted Self-Interviewing (CASI); Likert Scale; Question Order Effects; Response Latency; Semantic Differential Technique; Social Desirability

Further Readings

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.

Cacioppo, J. T., & Petty, R. E. (1979). Attitudes and cognitive response: An electrophysiological approach. *Journal of Personality and Social Psychology*, *37*, 2181–2199.

Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 247–282). Mahwah, NJ: Lawrence Erlbaum.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*, 3–25.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*, 46–54.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 5–53.

McClendon, M. J., & O'Brien, D. J. (1988). Question-order effects on the determinants of subjective well-being. *Public Opinion Quarterly*, *52*, 351–364.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal report on mental processes. *Psychological Review*, *84*, 231–259.
Osgood, C. E., Suci, G. J., & Tanenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois.
Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598–609.
Porier, G. W., & Lott, A. J. (1976). Galvanic skin responses and prejudice. *Journal of Personality and Social Psychology*, *33*, 317–326.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.
Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology*, *38*, 368–389.

Self-Selected Listener Opinion Poll (SLOP)

A self-selected listener opinion poll, also called SLOP, is an unscientific poll that is conducted by broadcast media (television stations and radio stations) to engage their audiences by providing them an opportunity to register their opinion about some topic that the station believes has current news value. Typically two local telephone numbers are broadcast for listeners to call to register their opinion on a topic. One number might be for those who agree with the issue, and the other might be for those who disagree. For example, a question might be to indicate whether a listener agrees or disagrees that "*the mayor should fire the police commissioner*."

Because the people who choose to call in do not represent a known target population, the findings from such polls have no external validity as they cannot be generalized to any particular population and therefore are not valid as measures of public opinion. Although these polls may provide some entertainment value for the station and its audience, especially for those who call in, they are not scientifically valid measures of news or public opinion.

Paul J. Lavrakas

See also Call-In Polls; Computerized-Response Audience Polling (CRAP); 800 Poll; External Validity; 900 Poll; Pseudo-Polls

Further Readings

Traugott, M. W., & Lavrakas, P. J. (2008). *The voter's guide to election polls* (4th ed.). Lanham, MD: Rowman & Littlefield.

Self-Selected Sample

A sample is self-selected when the inclusion or exclusion of sampling units is determined by whether the units themselves agree or decline to participate in the sample, either explicitly or implicitly.

How Does Self-Selection Enter Into a Sampling Design?

There are three main routes through which self-selection enters into a sampling design. The first is nearly ubiquitous, and the latter two are more preventable.

Refusals

When survey units are chosen by surveyors, but these units nonetheless elect not to participate (also called refusal-related nonresponse), self-selection occurs. Nonresponse can occur in probability or nonprobability sampling designs. If many such units elect not to participate, the representativeness of the resultant observed sample can be called into serious question as it may result in nonnegligible nonresponse bias. Self-selection can occur at the interview level (i.e., missing data due to a refusal to be interviewed) or at the item/question level (i.e., a refusal to answer a specific question or questions during the interview). For example, in a survey administered to a probability sample of adults in a particular city with the goal of estimating the city's mean income level, some persons will refuse to be interviewed, and of those who agree to be interviewed, some will refuse to report their income.

Volunteers

When survey units volunteer to be included in the sample, this introduces self-selection. Volunteer samples are one type of nonprobability sampling design. They are most common when rare, difficult-to-locate, demographic subpopulations are sampled, or when surveyors seek to obtain information from many people, quickly, at relatively low cost. For example, in an Internet survey of the prevalence of student drug use, students could be recruited by first asking school principals to volunteer their schools for participation, then asking teachers within schools to volunteer their classrooms to participate, and third asking students to volunteer to participate by going to a Web site to fill out the questionnaire.

Incidental Truncation

When units are sampled if, and only if, they engage in some behavior, then the distribution of the outcome variable is truncated at an unknown point. This is *incidental truncation*. For example, consider a survey in which researchers are interested in estimating the mean Scholastic Aptitude Test (SAT) score in each of the 50 United States. However, students' SAT scores can be sampled only if they choose to take the SAT, which results in incidental truncation because it is very likely that among those who do not take the SAT, there will be scores even lower (were they to take the test) than the lowest score among those taking the SAT.

What Is the Self-Selected Sample Problem?

The underlying statistical problem is that if the reasons for a sampling unit's selection into or out of the sample relate to the outcome variable of interest, then *self-selection bias* occurs. Quite simply, this bias occurs when self-selection is either a function of the survey unit's score on the outcome variable or is codetermined with the outcome variable. In this event, point estimates and standard errors will be biased, and statistical inference about population parameters from sample statistics may be invalid.

Returning to the first example given, this problem would occur if low-income persons refuse to report their income because they are too embarrassed. Returning to the second example, this problem would occur if the propensities for self-selection at each stage are (a) related to each other, and (b) based on unmeasured characteristics, such as motivation and obedience, which are, in turn, related to the outcome variable. Returning to the third example, one way this problem would occur is if students who are intending to go to college and have higher grades are more likely to take the SAT and also more likely to get higher SAT scores than would those who are not college bound were they to have taken the SAT. Should these biases be present, their severity would be positively related to the magnitude of the difference between those who participated and those who did not and also positively related to the proportion that elects not to participate.

There is one misconception involving self-selected samples that is worth noting. Researchers often assume that even if selection bias is present, estimates of the outcome of interest will be unbiased for the part of the outcome distribution not truncated by the self-selection process. That is, they assume they will be able to make valid inferences about the subpopulation of self-selecters but unable to generalize results to the larger population, which includes the noncompliers, nonresponders, and nonvolunteers. In fact, if selection bias is present, regression coefficients estimated for the subsample of self-selecters will be biased as well.

Solutions

Essentially, for unbiased point estimates and standard errors, researchers need to be able to assume that selection into or out of the sample is either random (i.e., a sampled unit is an independent draw from the marginal population distribution of the outcome variable) or conditionally random (i.e., a sampled unit is an independent draw from the conditional population distribution of the outcome variable, conditioning on measured covariates). In most instances, it is highly unlikely that sampling units self-select into or out of a sample for reasons completely independent of the outcome of interest. More often, sampling units can be thought of as self-selecting into or out of a sample for reasons that are conditionally independent from the outcome variable. If this assumption is met, several solutions are possible.

One solution that is available when dealing with a probability sample with nonresponse is to identify those units who opted not to participate, identify variables measured on both responders and nonresponders that are also related to the outcome variable, and stratify the sample on these measured variables. This amounts to constructing weighting classes. The surveyor can then adjust raw sampling weight for each unit by the inverse of their response propensity in the weighting class (i.e., weighting class adjustment).

Returning to the first example, final estimates could be re-weighted to give more weight to persons from strata with low probabilities of responding to the income question.

Another solution is to identify observable reasons why some sampling units chose not to participate and statistically control for these in a model-based analysis. This will yield a conditional estimate of the outcome of interest. Depending on the relations among the outcome, predictors, and selection variables, as well as the model employed, this can reduce bias. Returning to the third example, the survey researcher could conceivably measure—and covary out—potential selection variables such as high school grade point average and whether or not the student's school requires the SAT to be taken from a regression equation predicting SAT score from socioeconomic status. These selection covariates would be expected to relate to the outcome of interest, SAT score, as well as to other observed predictors in the model, such as socioeconomic status.

Many times, however, researchers simply are not able to find measured selection variables that explain the self-selection process, nor do they have a probability sampling design such that they can adjust sampling weights for nonresponse. This problem would occur if self-selection were based on the outcome variable itself or on unobservable variables correlated with the outcome. This situation is considerably more complicated.

One possible approach is to model the joint distribution of the outcome of interest, along with the distribution of the self-selection process. In a typical model-based analysis, the mechanism by which the outcome variable in the sampled population is generated is modeled in a hypothetical superpopulation. In this more complicated circumstance, however, the researcher also has to simultaneously model the self-selection mechanism by which the outcome variable in the sample is selected from the sampled population. Estimates from such joint models depend heavily on the plausibility of the underlying model assumptions for the sample selection model. If these are severely violated, estimates of the outcome of interest can be as biased as or more biased than ignoring the self-selection mechanism altogether.

Overall, it is of paramount importance for any presence of self-selection to be made transparent in the reporting of survey results so that the consumers of the research can be aware of hidden biases that may have affected results.

Sonya K. Sterba and E. Michael Foster

See also Convenience Sampling; Missing Data; Model-Based Estimation; Nonprobability Sampling; Nonresponse; Nonresponse Bias; Purposive Sample; Refusal; Representative Sample; Response Propensity; Self-Selection Bias; Superpopulation

Further Readings

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, *48*, 386–398.

Davidson, R., & MacKinnon, J. G. (2004). *Econometric theory and methods*. New York: Oxford University Press.

Wainer, H. (1986). *Drawing inferences from self-selected samples*. New York: Springer.

Self-Selection Bias

Self-selection bias is the problem that very often results when survey respondents are allowed to decide entirely for themselves whether or not they want to participate in a survey. To the extent that respondents' propensity for participating in the study is correlated with the substantive topic the researchers are trying to study, there will be self-selection bias in the resulting data. In most instances, self-selection will lead to biased data, as the respondents who choose to participate will not well represent the entire target population.

A key objective of doing surveys is to measure empirical regularities in a population by sampling a much smaller number of entities that represent the whole target population. Modern sampling theory is predicated on the notion that whether an entity is eligible for interview should be determined by a random mechanism as implemented by the researcher that ensures that, for defined subpopulations formed by a partition of the entire population, the probability of selection is either proportional to the number in the subpopulation or, after weighting, weighted sample size is proportional to the number in the subpopulation. Further, the notion that sampling is random rules out selection based on behaviors or attributes about which the researchers are attempting to learn. For example, if researchers seek to learn about political affiliation, the sample will be compromised if the probability of inclusion varies by the respondent's political affiliation. Unfortunately, virtually all survey samples of human beings are self-selected to some degree due to refusal-related nonresponse among the sampled elements. In some cases this merely contributes negligible bias, whereas in others the bias is considerable.

The problem with self-selected samples comes when a respondent chooses to do a survey for reasons that are systematically related to the behaviors or attributes under study. The starting point for the literature on *selectivity bias* dates back more than 30 years to the work of labor economists. Central to the selectivity bias literature is that the seriousness and intractability of the problem increase when selection into the sample is driven not by exogenous or predetermined variables (under the researcher's control) but by unmeasured effects that also influence the behaviors and other variables the survey researchers want to learn about. In the latter case, the threat to validity is large when the rate of nonresponse is also large. An all-volunteer sample is the worst case of nonresponse bias when no one is selected based upon a scientific sampling rule. Consequently, threats to validity peak with self-selected samples—a category into which, for example, far too many Internet polls fall. The goal of sampling is to reduce the scope for people to opt into a study based upon the measures under study. Thus, respondents should be chosen for a survey sample based upon some mechanism that is well understood and statistically independent of the researchers' measurement protocol.

When the respondent chooses the study rather than the study choosing the respondent, the respondent may opt into a study based upon predetermined, observable characteristics, such as age, race, sex, or region of origin or, more dangerously, based upon some characteristic that is respondent determined (or at least heavily influenced), such as political ideology, hours worked, religiosity, or other attitudes. When respondents choose a survey for reasons related only to their demographic characteristics, such as age, race, or sex, the damage to randomness often can be "undone" by judicious post-stratification weighting, so long as researchers know the correct universe estimates for these characteristics. However, when omitted variables affect both the propensity to volunteer and the measures under study, the situation becomes difficult, requiring substantial structure to undo the damage of a self-selected sample.

With political polling that aims to measure the population's ideology, the risk that some respondents might step forward to do the poll based on their ideology is a problem that cannot be undone by weighting, for if researchers already knew the breakdown of ideology in the population, why would they be conducting the survey in the first place? Thus, for example, there is a good reason organizations doing exit polls select every *n*th voter as opposed to all voters with particular bumper stickers.

Unfortunately, when it comes to sampling, the statistician proposes and the respondents dispose. Self-selection creeps in when a respondent's propensity to cooperate is related to the survey's measurement objectives. As cooperation rates fall for scientifically designed samples, the scope for self-selection increases. Consider an exit poll being done for a newspaper or television network that respondents broadly perceive as having a particular ideological predisposition (i.e., an organization that is viewed as very conservative or very liberal). If the interviewer reveals the name of the newspaper or network, respondents having the same worldview may be predisposed to cooperate, and others without that worldview may not.

Ironically, while academic survey organizations are frequently seen as more dispassionate collectors of data, they are frequently subject to regulation by institutional review boards that often circumscribe the ability of interviewers to secure cooperation from all respondents. With less latitude to convert reluctant respondents, completion rates decline, magnifying the impact of self-selection via differential cooperation.

Randall Olsen

See also Bias; Differential Nonresponse; External Validity; Institutional Review Board (IRB); Nonprobability Sample; Nonresponse Bias; Probability Sample; Response Propensity; Sampling Bias; Self-Selected Sample; Weighting

Further Readings

Gronau, R. (1974). Wage comparisons—a selectivity bias. *Journal of Political Economy*, *82*, 1119–1143.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited

dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492.
Olsen, R. (1977). *An econometric model of family labor supply*. Unpublished doctoral dissertation, University of Chicago.
Olsen, R. (1980). A least squares correction for selectivity bias. *Econometrica*, *48*, 1815–1820.

Semantic Differential Technique

The semantic differential measurement technique is a form of rating scale that is designed to identify the connotative meaning of objects, words, and concepts. The technique was created in the 1950s by psychologist Charles E. Osgood. The semantic differential technique measures an individual's unique, perceived meaning of an object, a word, or an individual.

The semantic differential can be thought of as a sequence of attitude scales. Using a 7-point bipolar rating scale, respondents are expected to rate an object. The 0 position typically means "neutral," 1 means "slightly," the 2 position means "quite," and 3 is "extremely." The scales are designed such that the left side is generally positive and the right is generally negative. This allows the semantic differential to measure intensity and directionality.

The rating scale consists of a list of bipolar responses. These responses are simply opposing adjectives. For example, the semantic differential might use the terms *rough* and *smooth* as its bipolar responses. Using an adapted Likert scale, the respondent chooses a point on the continuum to indicate to which term the object is most closely related. Once this has been completed, the researcher can "map" the respondent's connotations for the object. An example of a semantic differential is provided in Figure 1.

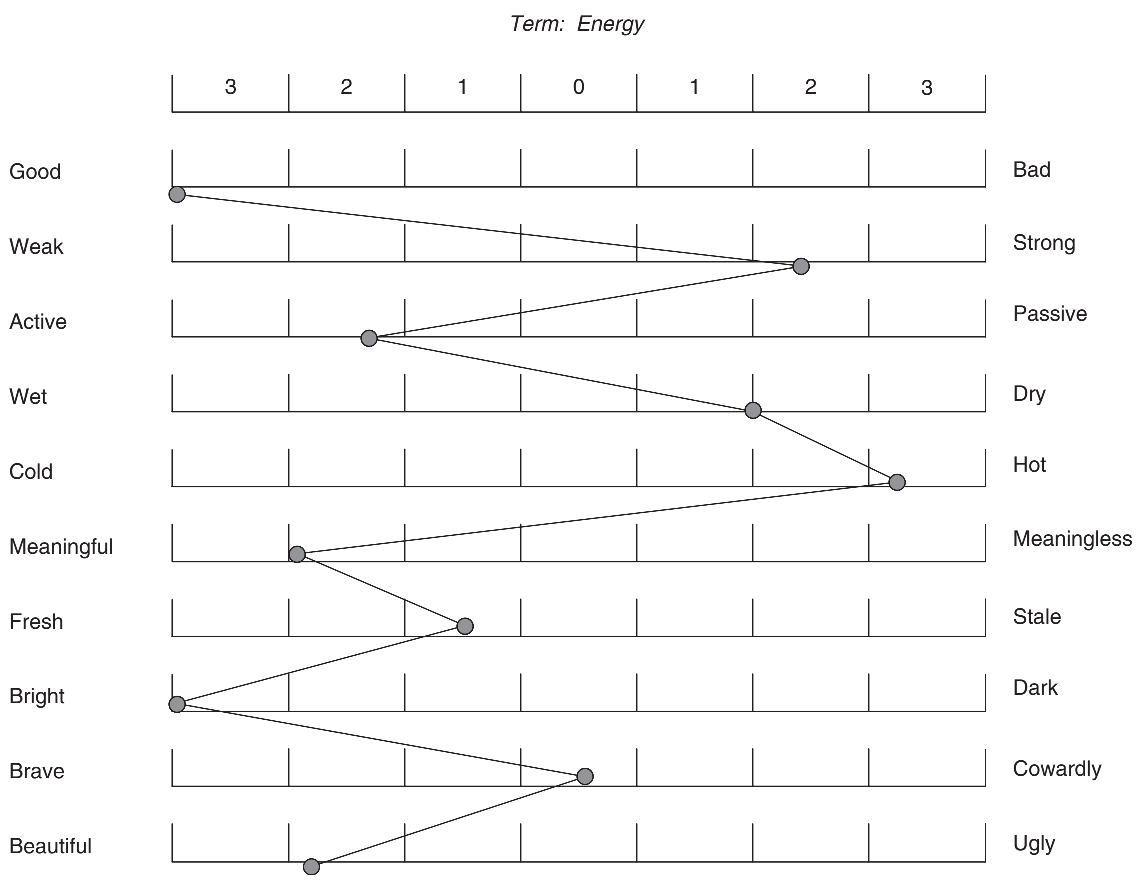

Figure 1 Example of the semantic differential technique

The Logic of the Semantic Differential Technique

The semantic differential is based on the following hypotheses:

1. The process of description or judgment can be conceived as the allocation of a concept to an experiential continuum, definable by a pair of polar terms.
2. Many different experiential continua, or ways in which meanings vary, are essentially equivalent and hence may be represented by a single dimension.
3. A limited number of such continua can be used to define a semantic space within which the meaning of any concept can be specified.

The first hypothesis assumes that discriminations in meaning cannot be finer or involve any more variables than are made possible by the sensory nervous system. The second assumes that the terms used are unique and independent of one another, which is why measurement is possible. The third hypothesis identifies a factor analysis form of methodology, and it allows for the opportunity of measuring meaning-in-general objectively.

Semantic Space

Factor analysis of the semantic differential data allows the researcher to explore a respondent's "semantic space." The semantic space represents the three underlying attitudinal dimensions that humans are hypothesized to use to evaluate everything. Research has demonstrated that these dimensions are present regardless of the social environment, language, or culture of the respondent. The three dimensions are *evaluation*, *power*, and *activity*.

The evaluation factor can be thought of as the good/bad factor. Common bipolar responses are "good/bad," "fresh/stale," "friendly/unfriendly," or "interesting/uninteresting." The power factor, which is sometimes called the potency factor, is the strong/weak factor. Common semantic differential responses for the power factor include "strong/weak," "powerful/powerless," "large/small," or "brave/cowardly." The activity factor is characterized as the active/passive factor. A number of bipolar responses can measure this, including "active/passive," "tense/relaxed," and "fast/slow."

Using these scales, the researcher can attain a reliable measure of a respondent's overall reaction to something. Researchers can obtain a subject's dimensional average by dividing the scales into their appropriate dimensions and averaging their response scores. Once completed, these measurements are thought of as the concept's profile.

Critique

Although the semantic differential technique has been widely used since its inception, there are some concerns about it. Theoretically, using the scale rests on the assumption that humans' connotations of a word are not the same, which is why this technique is needed. Paradoxically, the scale assumes that the chosen adjectives mean the same to everyone. For the scale to work, it must be assumed that humans share the same connotations for some words (i.e., the bipolar adjectives). This is a fairly strong assumption. For instance, looking at Table 1, the bipolar pairs of adjectives generally are set out with "positive" adjectives on the left. However, for the cold/hot dyad it is not clear that "cold" is always associated with positive thoughts. Indeed, depending upon the respondent's past, "cold" could easily evoke negative thoughts.

Another concern is that attitudes do not always match up with behavior. As such, attitudes can be poor predictors of action or behavior. The semantic differential ought to be able to overcome these concerns simply by design, but that does not mean that it can do so on its own. If respondents give socially desirable answers, it will negatively impact the reliability of the measure. Also, if a respondent begins to consistently answer in the same way (i.e., all neutral or always agreeing), the reliability must be questioned. Yet another critique is that the semantic differential does not actually identify individual emotions. Because of its design, the semantic differential technique cannot distinguish beyond the one continuum, but then it never was intended to do so in the first place.

James W. Stoutenborough

See also Attitudes; Bipolar Scale; Likert Scale; Rating; Respondent; Social Desirability

Further Readings

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Osgood, C. E., & Tzeng, O. (Eds.). (1990). *Language, meaning, and culture: The selected papers of C. E. Osgood*. Westport, CT: Praeger.
Snider, J. G., & Osgood, C. E. (Ed.). (1969). *Semantic differential technique: A sourcebook*. Chicago: Aldine.

SENSITIVE TOPICS

There is no widely accepted definition of the term *sensitive topics*, even though most survey researchers would probably agree that certain subjects, such as income, sex, and religion, are definitely examples of the concept. In their classic text *Asking Questions: A Practical Guide to Questionnaire Design*, Seymour Sudman and Norman Bradburn avoided the term altogether and instead talked about "threatening" questions.

Part of the problem is that topics or questions can be sensitive in at least three different, though related, senses. The first sense is that of *intrusiveness*. Some questions are inherently offensive to some (or most) respondents; some topics are seen as inappropriate for a survey. Respondents may find it offensive to be asked about their religion in a government survey or about their income in a study done by market researchers. Sudman and Bradburn used the phrase "taboo topics," and some topics are clearly out of bounds in certain contexts. It would be odd (and impolite) to ask a new coworker intimate details about his or her sexual life or medical history; in the same way, respondents may regard some survey topics or questions as none of the researcher's business. A second sense of "sensitivity" involves the risk that the information may fall into the wrong hands. Teenagers in a survey on smoking may worry that their parents will overhear their answers; respondents to the American Community Survey may worry that the Internal Revenue Service will be able to access their answers to the income questions on the survey. Questions are sensitive in this second sense when they raise concerns that some third party (whether another household member or some agency or business other than the survey firm) will learn what the respondents have reported. For example, in business surveys, responding firms may worry that their competitors will gain proprietary information about them. The final sense of sensitivity involves the social desirability of the behavior or attitude that is the subject of the question. A question is sensitive in this sense when it asks respondents to admit that their behavior has not lived up to some widely held standard or norm; such questions place the respondents at risk of embarrassing themselves. The embarrassing admission may involve perceived sins of commission (using illicit drugs or having had an abortion), or perceived sins of omission (not exercising enough or failing to vote in a recent election). Perhaps because the concept of sensitivity has multiple senses, there is no generally accepted method for measuring the sensitivity of a question or topic. Instead, most researchers rely on their professional judgment (including their intuitions) about which questions are likely to be sensitive.

These different senses of the term have somewhat different implications for surveys. For example, what makes a question intrusive appears to be the combination of the topic and sponsor of the survey. For example, although it would appear reasonable to most people for the Council of American Catholic Bishops to do a survey on religion, it may not appear reasonable to many for the U.S. Census Bureau to do so. Similarly, people worry more about disclosure risk or embarrassment when they have something they do not want others to find out and they fear the consequences of its becoming known. The consequences may, in turn, depend on who finds out. For a teenage girl, it is one thing for a classmate to learn that she occasionally smokes a cigarette but it may be quite a different thing for her parents to find out.

Consequences of Sensitivity

Despite respondents' potential objections to such topics, surveys often include questions about sensitive topics. To cite one example, since 1971 the federal government has sponsored a series of studies (most recently, the National Survey on Drug Use and Health) to estimate the prevalence and correlates of illicit drug use in the United States. Because surveys cannot always avoid asking about sensitive subjects, it is important to know how people react to them and how those reactions may affect the statistics derived from the surveys.

Asking about sensitive topics in surveys is thought to have three negative consequences for survey statistics. First, sensitive items may produce higher-than-normal levels of item nonresponse (missing data), reducing the number of cases available for analysis and possibly biasing the results. Often, the income question in a survey has the highest level of missing data of any item. The Current Population Survey imputes 20% or more of the responses to the income questions in that survey. These results suggest that one way respondents cope with sensitive questions is not to answer them.

A second way for respondents to avoid answering sensitive questions is not to take part in the survey at all. If the topic of the survey is known beforehand or if the respondent is worried that the survey information may be disclosed to third parties, he or she can refuse to participate. Like item nonresponse, unit nonresponse reduces sample sizes and can bias the results (e.g., if those with the most to hide are also the least likely to take part). Partly to counter such concerns, many federal surveys provide assurances that the data will not be disclosed to outsiders in an identifiable form. Research by Eleanor Singer and her colleagues suggests that these confidentiality assurances have a small but reliable positive effect on response rates.

Relatively little methodological work has explored the effects of question sensitivity on item and unit nonresponse, but a good deal has examined the third possible consequence of question sensitivity—respondents deliberately giving inaccurate answers to the questions, or misreporting. Methodological studies have examined overreports of socially desirable behaviors ranging from voting to attending church to exercising regularly. Similarly, they have explored underreporting of various undesirable behaviors, including having undergone an abortion, consuming too much fat in one's diet, and using illicit drugs. Still other studies have examined reports about sexual behaviors and attempted to explain why men report so many more sexual partners than do women. These studies tend to show high levels of systematic error (bias) in survey reports about sensitive behaviors. For example, more than 20% of nonvoters claim that they voted; similarly, surveys underestimate the number of abortions in the United States by 50% or more, presumably because the respondents underreport their abortions.

Coping With Sensitivity

Methodological textbooks recommend many tactics to reduce the effects of asking sensitive questions in surveys. Some of them attempt to address specific concerns about disclosure of the information to third parties. For example, as already noted, many surveys promise the respondents that the information they provide will remain confidential, partly to blunt any effects of sensitivity on unit nonresponse. Most of these recommended tactics attempt to improve the amount and accuracy of reporting. They include (a) increasing privacy, (b) asking indirect questions, (c) using "forgiving wording," and (d) other special question strategies and methods.

Increasing Privacy

Some methods used in surveys with sensitive questions increase the privacy of the data collection process in an attempt to reduce the motivation to misreport. For example, the data may be collected in a private setting, either away from anyone else in the respondent's home or in a setting outside the home. Evidence shows that interviews are often conducted in the presence of bystanders and, at least in some cases (e.g., when the bystander is the respondent's parent), this inhibits truthful reporting. Because of the threat that other family members may learn sensitive information, other settings, such as schools, may be better for collecting sensitive information from teenagers.

Another method that may increase the respondent's sense of privacy is self-administration of the questions. Both paper questionnaires and computerized self-administration (such as audio computer-assisted self-interviewing) increase reporting of sensitive information, and some studies indicate that computerized self-administration is even more effective than the self-administration via paper questionnaires. Collecting the data anonymously (with no identifying information that links the respondents to their answers) may produce further gains in reporting. However, having respondents put their completed questionnaires in a sealed ballot box does not appear to enhance the basic effect of self-administration, although there is little firm evidence on this point.

Indirect Question Strategies

A variation on the anonymity strategy is to use indirect (i.e., masked) questions that do not directly reveal the sensitive information to the interviewers or the researchers. One such set of methods is the *randomized response technique* (RRT), in which respondents use a randomizing device (a spinner or the flip of a coin) to determine the question they are supposed to answer. In one version of the technique, the respondent answers either the sensitive question (*Have you used cocaine in the last 30 days?*) or an unrelated innocuous question with a known probability of a "yes" answer (*Were you born in April?*). In another approach, the randomizing device determines whether the respondent is instructed to say "yes," say "no," or even answer the sensitive question. Although the randomized response technique seems to be effective (yielding more accurate answers), these gains come at a price. The estimates based on the RRT are more variable than estimates based on direct questions (because fewer respondents actually answer the sensitive question), and it is harder to determine the correlates of the sensitive behavior (because the researchers cannot be sure which individuals answered the sensitive question).

A strategy related to RRT is the *item count technique* (also called the *unmatched count technique* and the *list experiment technique*). This procedure involves asking respondents to say how many behaviors they have done on a list of behaviors; they are not asked to report which behaviors they did, just how many. Some respondents get a list that includes the sensitive item; the rest get the same list without that item. For example, the question might ask one group of respondents, *How many of the following have you done in the past 30 days: Bought new shoes, read a newspaper, donated blood, smoked marijuana, and visited a foreign country?* A second group gets the same list without the item on smoking marijuana. The evidence is mixed about how well the item count technique actually works and, like RRT, it has rarely been used in general population surveys.

Forgiving Wording and Other Wording Approaches

Survey textbooks often recommend "loading" sensitive questions to invite a socially undesirable answer. The question might presuppose the behavior (*How often have you smoked marijuana in the last month?*) or suggest that a particular "bad" behavior is very common (*We often find a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have the time*.). There is surprisingly little evidence one way or the other on the effectiveness of *forgiving wording*. Familiar wording (asking about *having sex* rather than *sexual intercourse*) does seem to increase reporting of sensitive information.

One other method has been used to reduce item nonresponse for sensitive questions. In collecting income or other financial information, researchers sometimes use an approach called *unfolding brackets*. For example, respondents who will not (or cannot) report an exact income figure get bracketing questions (*Was the amount more or less than $25,000? More or less than $100,000?*) that allow the respondents to be placed into a broad category. Some respondents are willing to answer the bracketing questions but not those calling for exact information; still, some refuse to provide either type of information.

Bogus Pipeline and Other Methods

Many studies show that an effective method for improving reports about sensitive topics is the *bogus pipeline*. This method involves leading respondents to believe that the interviewer or researcher can determine whether the respondent is telling the truth. Researchers have used a variety of means to convince the respondents that they can detect false reports, ranging from (bogus) polygraph-like devices to (real) biological assays that actually can detect false reports (such as urinalyses that can detect recent drug use). The bogus (or real) pipeline presumably reduces the respondent's motivation to misreport, since the truth will come out anyway. The bogus pipeline is not always very practical in a survey setting (how does one get the equipment to the respondent?) and, when the pipeline is bogus, researchers may be unwilling to engage in the requisite deception.

One other tactic is sometimes recommended—that of matching interviewers and respondents on background characteristics (such as sex). This seems to be based on the idea that respondents are more likely to confide in interviewers who are similar to them than to ones who are dissimilar, but there is little evidence on the effectiveness or ineffectiveness of interviewer–respondent matching. For example, a 1992 study by Paul J. Lavrakas found that men were twice as likely to report having sexually harassed someone when they were interviewed by a male interviewer than when they were interviewed by a female interviewer, and that women were three times as likely to report having sexually harassed someone when they were interviewed by a female interviewer than by a male interviewer.

Conclusion

Surveys that ask people sensitive questions are here to stay. Unfortunately, people come to surveys armed with a lifetime of experience of fending off unwelcome questions. They can avoid the questions by avoiding the survey entirely, by refusing to answer specific sensitive questions, or by deliberately misreporting. Many surveys have adopted self-administration to improve reporting; this and the randomized response technique both seem to be effective. The key to both procedures is that the interviewer is not aware of the embarrassing information being revealed. Unfortunately, none of the survey techniques for dealing with sensitive questions eliminates misreporting entirely. Research methodologists still have a lot to learn about how best to collect sensitive information in surveys.

Roger Tourangeau

See also Anonymity; Audio Computer-Assisted Self-Interviewing (ACASI); Confidentiality; List Experiment Technique; Misreporting; Missing Data; Mode of Data Collection; Overreporting; Privacy; Randomized Response; Refusal; Response Bias; Social Desirability; Underreporting; Unfolding Question; Unit Nonresponse

Further Readings

Barton, A. J. (1958). Asking the embarrassing question. *Public Opinion Quarterly*, *22*, 67–68.

Lavrakas, P. J. (1992, November). *Attitudes toward and experience with sexual harassment in the workplace*. Paper presented at the 14th annual conference of the Midwest Association for Public Opinion Research, Chicago.

Singer, E., von Thurn, D., & Miller, E. (1995). Confidentiality assurances and response: A quantitative review of the experimental literature. *Public Opinion Quarterly*, *59*, 66–77.

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Sequential Sampling

For survey sampling applications, the term *sequential sampling* describes any method of sampling that reads an ordered frame of N sampling units and selects the sample with specified probabilities or specified expectations. Sequential sampling methods are particularly well suited when applied with computers. They can also be applied for selecting samples of a population resulting from some other process: for example, cars coming off an assembly line, patients arriving at a clinic, or voters exiting the polls. Examples of sequential sampling schemes discussed in this entry include simple random sampling, systematic sampling, and probability proportional to size (PPS) sequential sampling.

Simple Random Sampling (Without Replacement)

Simple random sampling without replacement is defined as selecting one of the possible distinct samples of size n from a population of size N. There are $\binom{N}{n}$ such possible samples, and each has an equal probability of being selected. Other methods generally involve selecting random numbers between 1 and N, discarding any repeats, and retaining the first n distinct units selected. Random ordering before selecting a pre-specified chunk is often used in computerized selection of simple random samples. The sequential procedure requires selecting a random number, R_i, for each population element and comparing it to a conditional probability based on what has occurred up to this point. Select Unit 1 if $R_1 \le n/N$. If Unit 1 is selected, select Unit 2 if $R_2 \le (n-1)/(N-1)$; if Unit 1 is not selected, select Unit 2 if $R_2 \le n/(N-1)$. Proceed through the list decreasing the denominator for each new unit but decreasing the numerator only when a selection occurs.

Systematic Sampling

For the simplest case, where sampling is with equal probabilities and $k = N/n$ is an integer, a random integer, I, between 1 and k is drawn and when the Ith element is encountered it is included in the sample. I is then incremented by k, and the $(I+k)$th element is included in the sample when encountered. The process continues until n units have been designated for the sample.

A more general form of systematic sampling can be applied where sampling is with unequal probabilities and/or $k \neq N/n$. Define the desired probabilities of selection for each unit as π_i for $i - 1, 2, \ldots, N$. For an equal probability design, $\pi_i = n/N$. For unequal probability designs, it is only necessary that $0 < \pi_i$

≤ 1 and $\sum_{i=1}^{N} \pi_i = n$. To select the sample sequentially, it is necessary to draw a uniform (0,1) random number, R. For Unit 1, define $S = \pi_1$. If $R \leq S$, then select Unit 1 and increment R by 1. For subsequent Unit i increase S by π_i and if $R \leq S$, then select Unit i and increment R by 1.

PPS Sequential

PPS sequential sampling is defined for probability minimum replacement (PMR) sampling. Sampling without replacement is then shown as a special case. Rather than working with the probability of selection, PMR selection schemes work with the expected number of hits or selections for each unit designated by $E(n_i)$, where n_i is the number of times unit i is selected for a particular sample. When a size measure X_i is used, the expected number of selections per unit i is set as

$$E(n_i) = \frac{nX_i}{\sum_{i=1}^{N} X_i}.$$

Note that this formulation allows $E(n_i)$ to be greater than 1. The defining PMR principle is that for every sample, the actual value of n_i will be either the integer portion of $E(n_i)$ or one greater.

Application of PPS sequential sampling requires keeping track of the two sums during the process:

$$S_i = \sum_{j=1}^{i} E(n_j) \quad \text{and} \quad T_i = \sum_{j=1}^{i} n_j.$$

The first sum is partitioned into integer component, I_i, and a fractional component, F_i, with $0 \leq F_i < 1$. S_0 and T_0 are set to 0 to start the process. The sequential selection process proceeds as follows:

- If $T_{i-1} = I_{i-1}$ and $F_i = 0$ or $F_{i-1} \geq F_i > 0$, then $T_i = I_i + 1$ with probability 0.
- If $T_{i-1} = I_{i-1}$ and $F_i > F_{i-1} \geq 0$, then $T_i = I_i + 1$ with probability $\frac{F_i - F_{i-1}}{1 - F_{i-1}}$.
- If $T_{i-1} = I_{i-1} + 1$ and $F_i = 0$, then $T_i = I_i + 1$ with probability 0.
- If $T_{i-1} = I_{i-1} + 1$ and $F_i > F_{i-1} \geq 0$, then $T_i = I_i + 1$ with probability 1.
- If $T_{i-1} = I_{i-1} + 1$ and $F_{i-1} \geq F_i > 0$, then $T_i = I_i + 1$ with probability $\frac{F_i}{F_{i-1}}$.

The number of times a unit is selected is then computed as $n_i = T_i - T_{i-1}$.

An estimate of a population total is analogous to Horvitz-Thompson estimator for PPS without replacement sampling except selection probabilities, π_i, are replaced with expected sample sizes. In the variance formula and in the variance estimator, the pairwise probabilities π_{ij} are replaced by expectation of the product of achieved sample sizes for the two units, $E(n_i n_j)$. To allow for unbiased variance estimation, $E(n_i n_j)$ must be positive for all pairs of units. This can be achieved by first ordering the list along some meaningful stratification dimension and considering the ordering as a closed loop. Then a random starting unit is selected, and the process is applied for the complete ordered list. As an example, the sampling units on the frame (or within a single stratum) may be ordered along some continuation (such as income levels) within a geographic area. With two areas, the order of sorting on income may be specified as increasing in one area and decreasing in the other. Any two neighboring elements on the closed loop will have at least one characteristic in common: same area, similar income, or both. Similar sorting schemes can be set up for more ordering variables.

Note that if the expected sample sizes are all less than 1, the selection procedure produces a probability without replacement sample. If all expected sample sizes are equal and less than 1, then it produces an equal probability sample.

James R. Chromy

See also Probability Proportional to Size (PPS) Sampling; Sampling Without Replacement; Simple Random Sample; Systematic Sampling

Further Readings

Chromy, J. R. (1981). Variance estimators for a sequential selection procedure. In D. Krewski, R. Platek, & J. N. K. Rao (Eds.), *Current topics in survey sampling* (pp. 329–347). New York: Academic Press.

Fan, C. T., Muller, M. E., & Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, *57*, 387–402.

Sheatsley, Paul (1916–1989)

Paul B. Sheatsley was an early leader in the new field of survey research. He was born in New York City in 1916 and received a bachelor's degree from Princeton University in 1936. While following a career in journalism and public relations as editor of the *Boonton Tribune* in New Jersey (1937–1939) and public relations director for the Yankees' farm team, the Newark International Baseball Club (1939), Sheatsley started working as a part-time interviewer for George Gallup. Gallup then hired him to be field director of the Audience Research Institute (1940–1942). In early 1942, Harry Field, the founder of the recently organized National Opinion Research Center (NORC) at the University of Denver, approached his friend and former colleague, Gallup, and asked to "borrow" Sheatsley to head NORC's New York office and direct the important series of studies that NORC was to do for the Office of War Information (1942–1944). Once he had gone to NORC, Sheatsley never left. He headed the New York office until 1963 and directed many major studies, such as the foreign policy series for the U.S. State Department (1945–1957). Then he moved to NORC's headquarters at the University of Chicago to direct its new Survey Research Service. He led that department until 1976 and served as NORC's acting director in 1970–1971. From 1976 until his retirement in 1986, Sheatsley was a senior survey director heading such large-scale projects as the National Ambulatory Medical Care Survey. From 1986 until his death in 1989, he continued as a consultant to NORC.

Sheatsley made major contributions in the areas of survey professionalization, methodology, and substantive analysis. He attended the first conference of survey researchers ever held that was organized by Field in Central City, Colorado, in 1946. This conference led to a second conference in Williamstown, Massachusetts, in 1947 where the American Association for Public Opinion Research (AAPOR) was organized. Sheatsley became a leader of AAPOR. He served on its council longer than any other person, holding offices in each decade from the 1940s to the 1980s. He was AAPOR president in 1967–1968 and was awarded AAPOR's Exceptionally Distinguished Achievement Award in 1982. He was coeditor with Warren J. Mitofsky of *A Meeting Place: The History of the American Association for Public Opinion Research* (1992).

While directing many seminal surveys for NORC, Sheatsley conducted a long series of methodological studies to improve data quality. He published articles on such topics as interviewer training, question wording, the use of incentives, validation, and open- versus closed-ended items.

Substantively, Sheatsley did pioneering work in the area of intergroup relations. He was coauthor of a series of four articles published in *Scientific American* between 1956 and 1978 that has been recognized as the first trend studies on intergroup relations, and he studied the race riots in the 1960s. Another seminal study that he analyzed was the Kennedy Assassination Study in 1963 that measured the public response to the president's murder and served as the baseline for NORC's National Tragedy Study, conducted in the aftermath of the 9/11 terrorist attacks in 2001.

Sheatsley's liberal arts education and journalistic training served him well in the field of survey research. Besides mastering the complex task of managing large-scale, national surveys and the quantitative skill of data analysis, he was able to both write clear and precise questions easily understandable by both respondents and interviewers and then present results that were insightful, true to the data, and comprehensible to the reader.

Tom W. Smith

See also American Association for Public Opinion Research (AAPOR); George Gallup; National Opinion Research Center (NORC)

Show Card

Within the context of survey research, a show card or *show sheet* is a visual aid used predominantly during in-person surveys. It is a card, a piece of paper, or an electronic screen containing answer categories to a question, from which the respondent chooses the answer to the survey question. The respondent may either look at the answer categories listed on the show card when providing the answer or mark the answer directly on the show card. The answers listed on the show card can be in the form of words, numbers, scales, pictures, or other graphical representations.

Telephone interviews may also employ show cards, although mailing the show cards or setting up Internet sites with the digital equivalents of show cards can be logistically and financially impractical.

Show cards are used by survey organizations internationally. In the United States, the use of show cards is somewhat wide-scale, including the federal government and private sector organizations. However, because they usually require in-person administration, show cards are not as well known or as well researched as other survey tools. The primary purpose for using show cards is to reduce survey measurement error. Show cards reduce error by (a) encouraging respondents to provide more accurate answers by increasing the perceived confidentiality of the answers, and (b) making it easier for respondents to provide a more accurate answer, for example, through presenting the entire answer category set in one visual field.

Providing More Accurate Answers on Sensitive Issues

Respondents may be uncomfortable providing accurate responses to some questions. Sensitive information can be easier for respondents to provide if they can do so under an additional veil of confidentiality, even from the interviewer. Sensitive information can range from health information (anything from weight to sexual history) to demographic information (anything from age to income). Accepted thinking is that granting the respondent maximum privacy during data collection is conducive to obtaining accurate answers.

In some cases, the respondent himself or herself marks the correct answer on the show card without verbally stating to the interviewer what that answer is. A different approach is to precede the answer categories with letters or other abstract identifiers, as in Figure 1. This allows the respondent to provide abstract answers, such as letters or other abstract identifiers, instead of verbalizing the specifics of the answer (e.g., "less than $25,000").

Making It Easier for Respondents to Provide More Accurate Answers

Some questions may be difficult to answer because of the complexity of the response choices. Questions that contain lengthy or complex answer choices may use show cards so that the respondent can visualize and review the full field of possible answer categories before choosing the most accurate answer.

There is no standard for answer complexity level that benefits from the use of a show card. Some research organizations employ show cards for questions with as few as four or five answer categories.

SHOWCARD A

Please provide the letter (a, b, c, d, e, or f) next to the line that best matches your household's total income from last year. Your household's total income should include the money earned by all members of your household before taxes.

a) less than $25,000
b) $25,000 to $49,999
c) $50,000 to $74,999
d) $75,000 to $99,999
e) $100,000 to $150,000
f) more than $150,000

Figure 1 Income show card

Best Practices for the Use of Show Cards

Research on show cards is scarce at present. Thus, best practices for their use are experientially based rather than experimentally based. Some organizations appear to have their own idiosyncratic preferences for the frequency and appropriateness of show card use. A few general best practice principles detail commonalities shared by the majority of show cards.

As with other survey materials, researchers should aim for answer categories that are mutually exclusive and comprehensive. Because show cards offer visual representations of answer categories, the researchers need to consider the appearance of the card during its creation. There is some concern that show cards may be conducive to primacy effects (respondents choosing answers that disproportionately fall within the first few answer categories on the card), which should factor into the card design and pretesting considerations.

Show cards generally do not include "Refuse" or "Don't Know" as answer categories. The additional level of confidentiality afforded by the show card aims to reduce the need for someone to refuse, and the in-person nature of the data collection allows for follow-up probes by the interviewer in instances of "Don't Know" answers.

Agnieszka Flizik

See also Face-to-Face Interviewing; Field Survey; Graphical Language; Measurement Error; Primacy Effect; Privacy; Probing; Sensitive Topics

Further Readings

Davis, W., Wellens, T. R., & DeMaio, T. J. (1996). *Designing response scales in an applied setting*. Washington, DC: U.S. Census Bureau. Retrieved April 17, 2008, from http://www.census.gov/srd/papers/pdf/wd9601.pdf

Dillman, D. A. (2006). *Mail and internet surveys: The tailored method, 2007 update with new internet, visual, and mixed-mode guide*. Indianapolis, IN: Wiley.

General Registrar Office for Scotland. (2006). *Census test follow up survey: Information about Scotland's people*. Retrieved April 17, 2008, from http://www.gro-scotland.gov.uk/files1/the-census/2006-census-test-evaluation/j8566b.pdf

Research Triangle Institute. (2007). *2007 National Survey on Drug Use and Health—Showcard booklet*. Retrieved April 17, 2008, from http://www.oas.samhsa.gov/nsduh/2k7MRB/2k7Showcards.pdf

U.K. Data Archive. (2003). *Family Resources Survey: Show cards*. Retrieved April 17, 2008, from http://www.data-archive.ac.uk/doc/4803%5Cmrdoc%5Cpdf%5C4803userguide9.pdf

Significance Level

The significance level (also called Type I error rate or the level of statistical significance) refers to the probability of rejecting a null hypothesis that is in fact true. This quantity ranges from zero (0.0) to one (1.0) and is typically denoted by the Greek letter alpha (α). The significance level is sometimes referred to as the probability of obtaining a result by chance alone. As this quantity represents an "error rate," lower values are generally preferred. In the literature, nominal values of α generally range from 0.05 to 0.10. The significance level is also referred to as the "size of the test" in that the magnitude of the significance level determines the end points of the critical or rejection region for hypothesis tests. As such, in hypothesis testing, the *p*-value is often compared to the significance level in order to determine if a test result is "statistically significant." As a general rule, if the *p*-value is no larger than the significance level, the null hypothesis is rejected and the result is deemed statistically significant, thus supporting the alternative hypothesis.

The level of significance can refer to the Type I error for a single hypothesis test or for a family of simultaneous tests. In the latter case, the "experiment-wise" or "family-wise" significance level refers to the probability of making at least one Type I error over the collection of hypothesis tests that are contained in the family. So for example, a survey may contain questions to solicit data to be used to compare the average expenditures, household ownership percentage, and education levels across two possible geographic sectors of a particular county. Because there are three main variables of interest that are to be compared across the two geographical regions, the family-wise level of significance will refer to the probability of making a Type I error rate on at least one of the three hypothesis tests that are performed for this family of tests.

In another example of multiple comparisons, comparisons of the average length of unemployment were made across four racial post-strata after an omnibus analysis of variance (ANOVA) revealed that the average unemployment periods are not equal across the four race groups. One possible set of post-hoc multiple comparisons consists of all pairwise tests for differences in the average unemployment period for two race groups at a time (i.e., six pairwise tests).

A Bonferroni adjustment or other multiple comparisons adjustment is typically made to the overall nominal Type I error rate to ensure the proper significance level is achieved for the family of tests. For example, to ensure that the overall significance level for a family of three hypothesis tests, the nominal significance level, α, would be divided by 3, and the adjusted significance level, $\alpha/3$, would be used as the Type I error rate for each of the three hypothesis tests in the family. In the second scenario, six comparisons would be made, so the Bonferroni adjustment for multiple comparisons equates to using $\alpha/6$ as the Type I error rate for each of the pairwise comparisons.

Large values of α generally imply more powerful tests but also introduce a higher likelihood for rejecting a null hypothesis that is in fact true. Conversely, smaller α values imply less tolerance for making a Type I error; values of α that are too small make rejection of the null hypothesis virtually impossible, thereby reducing the statistical power of the test (i.e., increase Type II error). Thus, values for the significance level are generally fixed by the researcher prior to data collection (often at .05) to ensure proper protection against making a Type I error while allowing for a reasonable level of statistical power to be achieved for the particular hypothesis test or family of tests at hand.

Trent D. Buskirk

See also Alternative Hypothesis; Analysis of Variance (ANOVA); Null Hypothesis; *p*-Value; Statistical Power; Type I Error; Type II Error

Further Readings

Yates, D., Moore, D., & McCabe, G. (1999).*The practice of statistics*. New York: W. H. Freeman.

Simple Random Sample

There are two varieties of simple random samples: (1) with replacement and (2) without replacement. To draw a simple random sample, one must have a frame drawn up for the population of interest prior to sampling or at least know the size of the frame in advance. For a simple random sample without replacement, all $\binom{N}{n}$ possible subsets of n units have equal probability of selection. For a simple random sample with replacement, all N^n n-tuples of units have equal probability of selection. Although simple random samples are rarely used in sample surveys, they have played an important role in the development of the epistemological theory for survey research, and they continue to be useful in pedagogy, as an efficiency benchmark, and as a proving ground for demonstrating the properties of complex estimation strategies.

In most cases, a researcher who chooses a simple random sample design would be doing two things contrary to good judgment by ignoring all cost considerations and ignoring all prior knowledge about the population. Quite to the contrary, information about heteroscedasticity (i.e., differences in variances among variables) in the variable of interest or about its relationship to variables known for the entire population should be used to improve the efficiency of the sample through stratification, systematic selection, or differential probabilities of selection. Similarly, analysis of the components of cost often leads to the decision to draw a multistage sample. Usually, stratification, systematic selection, unequal probabilities of selection, and clustering are all employed in sample surveys. In fact, historians of survey methodology have dug hard to find examples where simple random samples were actually employed. Those, of course, could have been improved.

Regarding epistemology, simple random samples played an important role in Jerzy Neyman's communication of his theory of confidence intervals in his landmark 1934 paper. More recently, those in favor of likelihood-based inference (frequentist or Bayesian) have sometimes argued their points under the assumption that the design-based statistician would use a simple random sample. This is not exactly fair, perhaps, but the simplicity of the design does make it easier to grasp the fundamental differences in approaches. That same simplicity makes these designs useful in the teaching of survey methodology, as well as in demonstrating the asymptotic properties of complex estimation techniques (e.g., ratio-estimation and regression estimation, raking, and imputation). The "design effect" is a useful tool for comparing the efficiency of alternate designs and is defined with reference to the variance arising from a hypothetical simple random sample of the same size.

There are scattered earlier references to "simple random sampling," but the first formal use of the phrase appears to be Harald Cramer's 1946 text on mathematical statistics. The first sampling textbooks by Frank Yates in 1949 and by W. Edwards Deming in 1950 did not use the phrase at all. The next appearances were in the 1953 textbooks of William Cochran and of Morris Hansen, William Hurwitz, and William Madow. Cramer reserved it for sampling with replacement, whereas the others reserved it for sampling without replacement. Leslie Kish in his 1965 text used the term *simple random sampling* if without replacement and *unrestricted sampling* if with replacement, a term which Arthur Bowley and Jerzy Neyman used for both variants some 30 years earlier. Current usage is ambiguous with respect to replacement.

David Ross Judkins

See also Design Effects (*deff*); Frame; Multi-Stage Sample; Nonprobability Sampling; Random Sampling; Sampling Without Replacement; Stratified Sampling; Survey Costs

Further Readings

Kish, L. (2003). The hundred years' war of survey sampling. In S. Heeringa & G. Kalton (Eds.), *Selected papers*. Hoboken, NJ: Wiley-Interscience. (Reprinted from *Statistics in Transition*, 1995, *2*[5], 813–830)

Neyman, J. (1967). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. In *A selection of early statistical papers of J. Neyman*. Berkeley: University of California Press. (Reprinted from *Journal of the Royal Statistical Society*, 1934, *97*[4], 558–625)

Smith, T. M. F. (1994). Sample surveys 1975–1990; An age of reconciliation? *International Statistical Review*, *62*(1), 3–34.

Small Area Estimation

Survey data have been effectively used to provide suitable statistics for the target population and for many subpopulations, often called *domains* or *areas*. Domains may be geographical regions (e.g., states or counties), sociodemographic groups (e.g., nonwhite Hispanic women between 18 and 65 years) or other subpopulations. A domain or an area is considered "large" or "major" if the domain sample is sufficiently large so that it can provide a *direct estimate* of the domain parameter, for example, the mean, with adequate precision. A domain or an area is regarded as "small" if the domain-specific sample is not large enough to produce an estimate with reliable precision. Areas or domains with small samples are called *small areas, small domains, local areas, subdomains*, or *substates*.

Beginning in 1996, the U.S. Congress began to require that the Secretary of Commerce publish, at least biennially, current data related to incidence of poverty for states, counties, and local jurisdictions of government and school districts "to the extent feasible." State and county estimates of the number of 5- to 17-year-old children in poverty and those among 65 and older are required. Poverty estimates for children are used to allocate federal and state funds, federal funds nearly $100 billion annually in recent years. As such, small area estimation is very important for the well-being of many citizens.

A Brief Primer on Important Terms in Small Area Estimation

For m small areas, suppose $Y_{ij}, j = 1, N_i$ denote values of a response variable (Y) for the N_i units in the ith small area. Imagine one would like to estimate $\gamma_i = N_i - 1 \sum_{j=1}^{N_i} Y_{ij}$, the finite population mean. Suppose $\boldsymbol{X}$ is a vector of explanatory variables. If explanatory variables are available for all the sampled units in the ith small area, to be denoted for simplicity by $1, \ldots, n_i$, then a *unit-level model* is used. But if only direct estimates Y_i for γ_i and summary data x_i for explanatory variables are available at the small area level, then an *area-level model* is used. If indirect small area estimates are produced by fitting a model relating the response variable and explanatory variables, and prediction of a small area mean is obtained by substituting explanatory variables into the estimated model, one gets a *synthetic estimate*, denoted by $\hat{\gamma}$. Synthetic estimates are much too model dependent, susceptible to model failure, and not design-consistent. A composite estimate, which is a convex combination of Y_i and $\hat{\gamma}_{is}$, rectifies these deficiencies.

This entry considers only some of the basic aspects of small area estimation. For example, neither the time series and cross-sectional approach to small area estimation nor the interval estimation problem is considered here. For this and many other important topics, the advanced reader should consult J. N. K. Rao's *Small Area Estimation*.

Two Popular Small Area Models

Both linear and nonlinear models and both Bayesian and frequentist approaches are popular in small area estimation. While the *estimated best linear unbiased prediction (EBLUP)* approach is key to developing composite estimates based on mixed linear models, the empirical Bayes (EB) and the hierarchical Bayes (HB) approaches can be used for both linear and nonlinear models. Many model-based developments in small area estimation use normality assumptions. For normal linear models, the EBLUP and EB predictors of the small area means are identical.

The nested error regression model is a popular unit-level model given by

$$Y_{ij} = \boldsymbol{X}_{ij}^T\beta + v_i + e_{ij}, j = 1, \ldots, N_j, I = 1, \ldots, m, \quad (1)$$

where $\boldsymbol{X}_{ij}$ is a p-component vector of auxiliary variables, v_i and e_{ij} are independently distributed with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, $j = 1, N_j$, $i = 1, \ldots, m$. G. E. Battese and colleagues proposed this model to estimate areas under corn and soybeans for 12 counties in Iowa based on unit-level auxiliary data obtained from the LANDSAT satellite readings and unit-level response variable available from farm survey data.

An early application of the area-level model is by Robert E. Fay and R. A. Herriot in order to improve the direct estimator Y_i for estimating the per capita income of small places, denoted by μ_i. They assumed for their problem that a p-vector of auxiliary variables $\boldsymbol{x}_i$ was available for each area i given by county values, housing data from the census, and tax returns from the Internal Revenue Service. They assumed

$$Y_i = \mu_i + e_i, \quad \mu_i = x_i^T\beta + v_i, \quad i = 1, \ldots, m, \quad (2)$$

where $e_i \overset{ind}{\sim} N(0, D_i)$ and $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$. Sampling variances D_is are assumed to be known.

For brevity, this discussion is confined to unit-level models under the normality assumption. The next two sections discuss, respectively, the EBLUP (or EB) and HB approaches to small area estimation.

EBLUP and Mean Squared Error

This section focuses on the nested error regression model. Rao discusses general mixed linear models. For large values of N_i, γ_i can be approximated by $\mu_i = \bar{\boldsymbol{X}}_i^T\beta + v_i$, where $\bar{\boldsymbol{X}}_i = N_i^{-1}\sum_{j=1}^{N_i}\boldsymbol{X}_{ij}$, and we use EBLUP of μ_i to predict γ_i. Let $\boldsymbol{Y}^{(1)} = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{m_1}, \ldots, Y_{mn_m})^T, \boldsymbol{X}^{(1)} = (\boldsymbol{X}_{11}, \boldsymbol{X}_{1n_1}, \ldots, \boldsymbol{X}_{m1}, \ldots, \boldsymbol{X}_{mn_m})^T$, $\boldsymbol{\Sigma}_{11} = Diag(\sigma_\epsilon^2\boldsymbol{I}_{n_1} + \sigma_v^2 11^T, \ldots, \sigma_\epsilon^2\boldsymbol{I}_{n_m} + \sigma_v^2 11^T), \boldsymbol{I}_p$ is a p × p identity matrix and $\mathbf{1}$ is a vector of ones. For known variance components $\psi = (\sigma_v^2, \sigma_\epsilon^2)^T$, the best linear unbiased prediction (BLUP) of μ_i is

$$\tilde{\mu}_i(\psi) = \bar{\boldsymbol{X}}_i^T\tilde{\beta}(\psi) + \delta_i(\bar{Y}_i s - \bar{\boldsymbol{x}}_{is}^T\tilde{\beta}(\psi)), \quad (3)$$

where $\bar{Y}_{is}$ is the ith area sample mean. Similarly, $\bar{\boldsymbol{x}}_{is}$ is defined. Also, $\tilde{\beta}(\psi)$ is the weighted least squares estimator of β based on Equation 1 corresponding to the sampled units, and $\delta_i = \sigma_v^2(\sigma_v^2 + \sigma_\epsilon^2 n_i^{-1})^{-1}$. For unknown variance components ψ, the BLUP cannot be calculated. Let $\hat{\psi}$ denote an estimator of ψ, estimated by ANOVA methods or maximum likelihood or residual maximum likelihood methods. For subsequent discussion, let us assume ANOVA estimates of the variance components are used. Then $\tilde{\mu}_i(\hat{\psi})$ is an EBLUP of μ_i.

The mean squared error (MSE) of an EBLUP of μ_i, given by $E[\tilde{\mu}_i(\hat{\psi}) - \mu_i]^2$, usually has no closed-form expression. A second-order approximation to this MSE ignoring all terms of order lower than $1/m$ is studied by Prasad and Rao and others. Indeed the approximation

$$MSE = E[\tilde{\mu}_i(\hat{\psi}) - \mu_i]^2 = g_{1i}(\psi) + g_{2i}(\psi) + g_{3i}(\psi),$$

where $g_{1i}(\psi) = (1 - \delta_i)\sigma_v^2$, and

$$g_2 i(\psi) = (\bar{\boldsymbol{X}}_i - \delta_i\bar{\boldsymbol{x}}_{is})^T(\bar{\boldsymbol{X}}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}(\psi)X^{(1)})^{-1}(\bar{\boldsymbol{X}}_i - \delta_i\bar{\boldsymbol{x}}_{is}), g_{3i}(\psi) = \frac{var(\sigma_v^2\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2\hat{\sigma}_v^2)}{n_i^2(\sigma_v^2 + \sigma_\epsilon^2/n_i)^3}. \quad (4)$$

Note that $g_{1i}(\psi) = 0(1)$ but $g_{2i}(\psi), g_{3i}(\psi)$ are $0(m^{-1})$. An estimator MSE is second-order unbiased if $E[mse] = MSE + 0(m^{-1})$. According to N. G. N. Prasad and Rao, $mse = g_{1i}(\hat{\psi}) + g_{2i}(\hat{\psi}) + 2g_{3i}(\hat{\psi})$ is second-order unbiased.

Hierarchical Bayes Approach to Small Area Estimation

For the nested error regression model, Gauri Datta and Malay Ghosh considered the HB model

1. $Y_{ij}|\beta, v_1, \ldots, v_m, \sigma_v^2, \sigma_\epsilon^2 \sim N(X_{ij}^T\beta + v_i, \sigma_\epsilon^2)$ independently $j = 1, \ldots, N_i, I = 1, \ldots, m$;
2. $v_i|\beta, \sigma_v^2, \sigma_\epsilon^2 \sim N(0, \sigma_\epsilon^2)$, independently $I = 1, \ldots, m$;
3. Independent prior on the second-stage parameters: $\boldsymbol{\beta}$ an improper uniform prior on R^p, σ_v^2 and $\sigma_\epsilon^2 \sim$ inverse gamma prior (possibly improper).

The HB predictor for γ_i is obtained from the predictive distribution of the unsampled units $Y_{ij}, j = n_i + 1, \ldots, N_i, i = 1, \ldots, m$ given the sampled units. Datta and Ghosh gave the predictive distribution in two steps. In Step 1, the predictive distribution is given for fixed hyperparameters σ_v^2 and σ_ϵ^2. In Step 2, the posterior distribution of the hyperparameters is given. Under squared error loss, for known hyperparameters, the Bayes predictor is also the BLUP. The HB predictor, typically with no closed-form, is obtained by integrating the above Bayes predictor with respect to the posterior distribution of the hyperparameters. Instead of assigning priors to the hyperparameters, if one estimates them from the marginal distribution of the data and replaces the variance components by their estimates in the Bayes predictor of γ_i, the result is the EB predictor. In fact, the EB predictor is identical to the EBLUP of γ_i. The HB predictor and associated measure of uncertainty given by the posterior variance can be computed by numerical integration or *Gibbs sampling*.

While the EBLUP is applicable to mixed linear models, the HB and the EB approaches can be applied even to generalized linear models, thereby making a unified analysis of both discrete and continuous data feasible.

Gauri Sankar Datta and Malay Ghosh

See also Composite Estimation; Parameter

Further Readings

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, *83*, 28–36.

Datta, G. S., & Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics*, *19*, 1748–1770.

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, *10*, 613–627.

Datta, G. S., Rao, J. N. K., & Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, *92*, 183–196.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistical Association*, *74*, 269–277.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science*, *9*, 55–93 (with discussion).

Lahiri, P., & Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, *90*, 758–766.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, *85*, 163–171.

Rao, J. N. K. (2003). *Small area estimation*. Hoboken: NJ: Wiley.

Snowball Sampling

Snowball sampling is a technique that can be applied in two survey contexts. The first context involves surveying members of a rare population. The second involves studying mutual relationships among population members. In both cases, respondents are expected to know about the identity of other members of the same population group.

Studying Rare Populations

In this context, snowball sampling is a nonprobability sampling technique. The general objective is to identify members of the rare population. It involves identifying one or more members of a rare population and asking them to name other members of the same population. These additional persons are then contacted and asked to name additional persons in the rare population; and so forth. The process continues until an adequate sample size has been obtained or until no new names are elicited from the process.

If terminated when adequate sample size is obtained, the method yields a sample, but not a probability sample.

If the population can be restricted in some way, say to a limited geographic area such as a county, snowball sampling may be successful as a rare population frame-building technique. To be successful, several rounds of the process must be conducted, and the initial sample should be large and adequately distributed among the rare population members. Within this restricted population, the identified rare population members can then be sampled using probability sampling techniques, or a complete enumeration (census) may be conducted. If the limited geographic areas are first-stage units in a multi-stage probability sample design, the approach can yield an estimate for a larger target population.

If some members of the targeted rare population are isolated from the remainder of the population, they are not likely to be named even after several rounds of enumeration. Serious coverage problems may remain even if the process is carried out diligently.

Studying Relationships

In the early 1960s, sociologist Leo Goodman proposed a probability sample–based method for studying relationships among individuals in a population. An initial zero-stage (Stage 0) probability sample is drawn. Each person in the sample is asked to name k persons with some particular relationship; example relationships are best friends, most frequent business associate, persons with most valued opinions, and so on. At Stage 1 these k persons are contacted and asked to name k persons with the same relationship. The Stage 2 sample consists of new persons named at Stage 1, that is, persons not in the original sample. At each subsequent stage, only newly identified persons are sampled at the next stage. The process may be continued for any number of stages, designated by s.

The simplest relationships involve two persons where each names the other. If the initial sample is a probability sample, an unbiased estimate of the number of pairs in the population that would name each other can be obtained. More complex relationships such as "closed rings" can be studied with more stages of sampling. For example, person A identifies person B; person B identifies person C; and person C identifies person A.

If the initial sample is drawn using binomial sampling so that each person has probability p of being in the sample and $s = k = 1$, an unbiased estimate of the number of mutual relationships in the population designated by M_{11} is

$$\hat{M}_{11} = \frac{y}{2p},$$

where y is the number of persons in the Stage 0 sample who named a person who also names them when questioned either in the initial sample or in Stage 1.

The theory for estimating the population size for various types of interpersonal relationships has been, or can be, developed assuming binomial sampling and may apply, at least approximately, when using other initial sample designs more commonly applied in practice, for example, simple random sampling (without replacement).

James R. Chromy

See also Coverage Error; Multi-Stage Sample; Nonprobability Sampling; Probability Sample; Rare Populations; Respondent-Driven Sampling (RDS); Sampling Frame; Sampling Without Replacement

Further Readings

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, *32*, 148–170.

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, *149*, 65–82.

Social Capital

Building on the work of sociologist James Coleman, political scientist Robert Putnam popularized the term *social capital* to describe how basic features of civic life, such as trust in others and membership in groups, provides the basis for people to engage in collective action. Even though social capital is not explicitly political, it structures various types of activities that are essential to maintaining civil and democratic institutions. Thus, *social capital* is defined as the resources of information, norms, and social relations embedded in communities that enable people to coordinate collective action and to achieve common goals.

It is important to recognize that social capital involves both psychological (e.g., trusting attitudes) and sociological (e.g., group membership) factors and, as such, is a multi-level construct. At the macro level, it is manifested in terms of connections between local organizations, both public and private. At the meso level, it is observed in the sets of interpersonal networks of social affiliation and communication in which individuals are embedded. And at the micro level, it can be seen in the individual characteristics that make citizens more likely to participate in community life, such as norms of reciprocity and feelings of trust in fellow citizens and social institutions.

Research on social capital, despite its multi-level conception, has focused on the micro level with individuals as the unit of analysis, typically using cross-sectional surveys to measure citizens' motivation, attitudes, resources, and knowledge that contribute to the observable manifestation of social capital: civic participation. The meso-network level is represented through individuals' reports of their egocentric networks in terms of size and heterogeneity as well as frequency of communication within these networks. Examinations of individuals' connections to community institutions and the connections among them are rare. These studies have been restricted to examining individuals' perceptions and attitudes regarding specific local institutions and the community generally (e.g., community attachment) as they relate to participation.

Most prominent among these institutional assessments has been political trust or trust in government, again measured mainly through individual-level survey assessments. Trust developed in interactions with social groups and government institutions is thought to function as a heuristic that is applied to decisions to participate in collective action efforts and is seen as foundational to the decision to become involved in civic life. The experience of participating in community projects, volunteering, and engaging in other membership activities reinforces feelings of trust and norms of cooperation, encouraging future civic involvement.

Survey measurement of civic participation, discussion networks, and social trust have often centered on the relationship between these indicators of social capital and patterns of media use. Survey evidence largely confirms that participation and trust have slipped in tandem, contributing to the erosion of community life. Changes in media adoption and use—for example, rising rates of television usage and declines in newspaper readership—across generational cohorts is thought to explain this decline, with television use both privatizing leisure time and presenting an increasingly harsh picture of the social world in televised representations of social reality. The combination was theorized to explain the correspondence between the rise in television use and the decline in social capital. Recent survey evidence from Dhavan Shah and his colleagues, from both cross-sectional assessments and panel survey designs, calls these assumptions into question. Instead, this research finds viewing news, documentary, and dramatic content can have pro-civic effects. This logic has been extended to the Internet, which has also been found to sustain social capital when used to gather information and strengthen social linkages.

Dhavan V. Shah and
Homero Gil de Zuniga

See also Trust in Government

Further Readings

Brehm, J., & Rahn, W. M. (1997). Individual level evidence for the causes and consequences of social capital. *American Journal of Political Science*, *41*, 999–1023.

Coleman, J. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.

Putnam, R. D. (1993). *Making democracy work: Civic traditions in modern Italy*. Princeton, NJ: Princeton University Press.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, *6*, 65–78.

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Shah, D. V., Cho, J., Eveland, W. P., Jr., & Kwak, N. (2005). Information and expression in a digital age: Modeling Internet effects on civic participation. *Communication Research*, *32*(5), 531–565.

Shah, D. V., McLeod, J. M., & Yoon, S. H. (2001). Communication, context and community: An exploration of print, broadcast and Internet influences. *Communication Research*, *28*, 464–506.

Social Desirability

Social desirability is the tendency of some respondents to report an answer in a way they deem to be more socially acceptable than would be their "true" answer. They do this to project a favorable image of themselves and to avoid receiving negative evaluations. The outcome of the strategy is overreporting of socially desirable behaviors or attitudes and underreporting of socially undesirable behaviors or attitudes. Social desirability is classified as one of the respondent-related sources of error (bias).

Social desirability bias intervenes in the last stage of the response process when the response is communicated to the researcher. In this step, a more or less deliberate editing of the response shifts the answer in the direction the respondent feels is more socially acceptable. Since the beginning of survey research, there have been many examples of socially desirable answers: for example, overreporting of having a library card, having voted, and attending church and underreporting of bankruptcy, drunken driving, illegal drug use, and negative racial attitudes.

The concept of social desirability has four nested characteristics: (1) The highest layer is a cultural characteristic, followed by (2) a personality characteristic, (3) mode of data collection, and (4) an item characteristic. The cultural characteristic is determined by the norms of that particular group or culture. For example, social desirability differs in conformist and individualist societies. Members of individualist societies tend to reveal more information about themselves to out-group representatives (the interviewer/researcher) than members of collectivist societies where the distinction between in-group and out-group is sharper. Within a society, specific cultural groups differ in level of social desirability. For example, studies conducted in the United States have shown higher scores of social desirability for minority groups when compared to majority groups.

The second characteristic is tied to a personality trait. Researchers describe it as the need to conform to social standards and ultimately as a response style. Some scales have been developed to measure the tendency of respondents to portray themselves in a favorable light, for example, the Marlowe-Crowne scale, the Edwards Social Desirability scale, and the Eysenck Lie scale.

The third characteristic is at the mode of data collection level. Social desirability has been found to interact with some attributes of the interviewer and the respondent, such as race/ethnicity, gender, social class, and age. One of the most consistent findings in the literature is that self-administrated methods of data collection, such as mail surveys and Internet surveys, decrease the prevalence of social desirability bias. The general explanation is that the absence of the interviewer reduces the fear of receiving a negative evaluation, and responses are, therefore, more accurate. The item characteristic is the question wording. For items that have shown social desirability bias or that are expected to show it, some question wording techniques successfully reduce it. Methods include loading the question with reasonable excuses, using forgiving words, and assuming that respondents have engaged in the behavior instead of asking if they have. Another strategy is the randomized response method.

From a practical point of view, the researcher should be aware of potential social desirability effects—especially in cross-cultural research. Although the researcher has no control over the cultural and personal characteristics, question wording and mode of data collection can be used to decrease potentially socially desirable responses. Particular care should be taken when mixing modes of data collection.

Mario Callegaro

See also Cognitive Aspects of Survey Methodology (CASM); Interviewer Effects; Mode Effects; Mode of Data Collection; Randomized Response; Respondent-Related Error; Sensitive Topics

Further Readings

De Maio, T. J. (1984). Social desirability in survey measurement: A review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 257–282). New York: Russell Sage.

Johnson, T. P., & Van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–204). Hoboken, NJ: Wiley.

Krysan, M., & Couper, M. P. (2003). Race in the live and the virtual interview: Racial deference, social desirability, and activation effects in attitude surveys. *Social Psychology Quarterly*, *66*, 364–383.

Social Exchange Theory

Social exchange is the theoretical approach most frequently invoked by survey methodologists when interpreting the decisions people make about whether or not to participate in surveys. In essence, social exchange theory holds that many interactions proceed from the assessment of costs versus benefits. Exchanges occur in purest form in economic transactions, in which it is fully realized that the exchange of goods or services for money or for barter is rational and voluntary, with the respective values for each party understood. Because money is so liquid, little social relationship is required for economic exchanges, and the individual is as often exchanging with the market as with other individuals. Social exchange theory widens the focus to a broad social realm, which includes intangibles such as maintenance of tradition, conformity to group norms, and self-esteem. The nature of the relationship between those exchanging then becomes important.

Examples of a social exchange explanation of seemingly irrational behavior appear in print within long-ago anthropological accounts of tribal life, for example, Bronislaw Malinowski's 1920s analysis of Kula exchange (the circulation of necklaces and armlets among Trobriand Islanders). Social exchange became a more widely known school of thought in work in the United States with the late 1950s and 1960s contributions by figures such as George C. Homans, J. W. Thibault and Harold H. Kelly, and Peter Blau.

By the mid-1970s, explicit applications were appearing of social exchange ideas to the understanding of nonresponse on surveys. An important statement was in *Mail and Telephone Surveys: The Total Design Method for Surveys* by sociologist Don Dillman in the late 1970s. Part of the total design was attending to several details of survey fieldwork likely to lower the cost and raise the rewards of participation for any sampled person. A stamped and addressed return envelope, for example, would lower the cost of responding to a mailed questionnaire, not just financially to a slight degree, but with a saving in time. Hand-signed cover letters on a questionnaire were a reward that acknowledged the importance of each individual drawn for the sample. Printing questionnaires in booklet form (in the pre–desktop publishing days of Gestetner machines) reduced the eyestrain costs of filling in a questionnaire while underscoring the importance of the research topic. Cash incentives drew their effectiveness primarily from social rather than economic exchange theory, when presented in prepaid form in small denominations. The money was symbolic, with high cultural resonance as a gift. Each item within the total design might seem slight, but the cumulative effect led to some highly respectable response rates on mailed surveys.

There is, however, a danger of overapplying social exchange theory as the explanation for why people do or do not respond to survey requests. Robert M. Groves and his colleagues have raised awareness of other factors not fitting so centrally into the social exchange framework. Especially on a telephone survey using unannounced calls, time does not exist for social exchange calculations to take place. Social exchange notions are most clearly valid where some degree of relationship exists between people, even if the relationship arises simply from the sequence of postal contacts in a mailed survey. When the telephone rings unexpectedly, and especially in a world of frequent beseeching telephone contacts from telemarketers and charities, most of the mental effort goes into making an instantaneous decision of whether and/or how to say "No." The decision making becomes heuristic, based on previous mental associations and rules of thumb. A "sing-song" voice, for example, instantly telegraphs that a call is not from a personal acquaintance.

Social exchange theory is a broad perspective that can embrace a wide conceptual territory. It is not so much that a cognitive heuristic interpretation of survey response decisions is diametrically in opposition to an exchange perspective as that heuristics are a particularly behaviorist version of exchange. The depth of decision making is much shallower on a cold telephone call than a decision, for example, about whether to continue participation in a panel study in which several household interviews have already taken place.

Thinking about the kind of social exchange that takes place around response decisions sensitizes the survey researcher to the kinds of people who are being included and excluded from their survey net. At the level of aggregates, exchange theory can guide researchers toward effective techniques, as Dillman's work demonstrated. For those seeking exact propositions about who among the public will and will not consent to a particular survey, social exchange theory is bound to disappoint. For unlike more behaviorist approaches, a prediction from social exchange theory requires full knowledge of a person's attitude toward survey research, the topic being examined, and whether the field contacts create a sense of benefit or at least minimize cost. If such information were already in hand from a sampling frame, a survey would hardly be necessary. It is no coincidence that modern empirical work on social exchange theory within social psychology involves laboratory settings where full background information indeed can be assembled.

John Goyder and Luc Boyer

See also Economic Exchange Theory; Nonresponse; Total Design Method (TDM)

Further Readings

Cropanzano, R., & Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. *Journal of Management*, *31*, 874–900.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method for surveys*. New York: Wiley.

Goyder, J., Boyer, L., & Martinelli, G. (2006). Integrating exchange and heuristic theories of

survey nonresponse. *Bulletin de Méthodologie Sociologique*, *92*, 28–44.
Homans, G. C. (1958). Social behavior as exchange. *American Journal of Sociology*, *62*, 597–606.

Social Isolation

Theories of social isolation (or "social location") have been used to explain lower cooperation in responding to surveys among certain subgroups in society, such as the elderly, minority racial and ethnic groups, and lower socioeconomic status groups. The social isolation theory of unit nonresponse states that subgroups in a society that are less connected to the dominant culture of the society—that is, those who do not feel part of the larger society or bound by its norms—will be less likely to cooperate with a survey request that represents the interests of the dominant society. According to the leverage-saliency theory, respondents decide to participate in a survey depending on survey attributes such as how long the interview might take, the presence of an incentive, and what the data might be used for. They may also make this decision depending on who is sponsoring the survey and what the topic of the survey is. The hypothesis of social isolation comes into play in these last two aspects.

Survey researchers have noted that sometimes sentiments of "civic duty" prompt survey participation and that these feelings of civic duty are associated with survey participation, especially when a properly constituted authority is requesting the participation of the sampled respondent. In theory, organizations with greater legitimacy, for example, those representing federal government agencies, are more likely to positively influence a respondent's decision to participate in a survey than, for example, a commercial survey research firm with less authority. However, according to some researchers, individuals who are alienated or isolated from the broader society feel less attachment to that society and thus have a lower sense of civic duty, which in turn will lead to a higher refusal rate in surveys conducted by and for those perceived as being in authority. Similar reasoning is often advanced regarding interest in and knowledge of the survey topic, in particular when focusing on political and election surveys. There are a number of studies that link political participation with survey participation. Results from these studies indicate that individuals who feel more isolated from the dominant society may also lack interest in politics and may feel less willing to respond to a political poll. In addition, nonresponse among members of socially isolated groups may be due to *self-disqualification*. That is, socially isolated individuals may feel they are not qualified to give an opinion on the topic or that their opinion is not valued.

Aspects of social isolation are both structural and social-psychological. For this reason, multiple indicators of social isolation should be used to study its relationship with survey cooperation. With demographic indicators (e.g., age) as a proxy for social isolation, some studies have found that elderly people are less likely to cooperate than are other subgroups. Elderly people often have fewer ongoing relationships with the larger society and thus are more isolated from the larger society. However, the effect of older age also has been found to be mitigated by living conditions and household composition. Single-person households tend to be less cooperative, whereas households with young children tend to be more cooperative. It is hypothesized that parents of young children are more likely to interact with strangers in the community on a daily basis, thereby increasing the level of connection with the society and thus also their likelihood of cooperating in surveys when a stranger (the interviewer) contacts them to seek cooperation with the survey.

If a survey is conducted using a sample that previously responded to an earlier wave of the survey, the data collected from the previous wave can be used to compare respondents and nonrespondents to the subsequent survey in terms of more direct indicators of social isolation, including variables such as time spent on voluntary work or contacts with neighbors, if those types of variable were measured in the previous wave. Studies of nonrespondents have found, for example, that people who spend more time on voluntary work tend to be more cooperative with survey requests, while those who have no close relatives or have no contact with neighbors are less likely to cooperate.

Mario Callegaro and Femke De Keulenaer

See also Cooperation; Leverage-Saliency Theory; Nonresponse; Refusal; Unit Nonresponse

Further Readings

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Singer, E. & Corning, A. (2000). Leverage-salience theory of survey participation. *Public Opinion Quarterly*, *64*, 299–308.

Loosveldt, G., & Carton, A. (2002). Utilitarian individualism and panel nonresponse. *International Journal of Public Opinion Research*, *14*, 428–438.

Voogt, R. J. J., & Saris, W. E. (2003). To participate or not to participate: The link between survey participation, electoral participation, and political interest. *Political Analysis*, *11*, 164–179.

Solomon Four-Group Design

The Solomon four-group design is an experimental design that assesses the plausibility of *pretest sensitization effects*, that is, whether the mere act of taking a pretest influences scores on subsequent administrations of the test. For example, if respondents complete a questionnaire measuring their knowledge of science as a pretest, they might then decide to subsequently seek answers to a few unfamiliar equations. At the posttest they might then score better on the science test compared to how they would have scored without taking the pretest. Meta-analytic results suggest that pretest sensitization does occur, although it is more prevalent for some measures than others, and the more time passes between pretest and posttest, the less likely a testing effect will occur.

In the Solomon four-group design, the researcher randomly assigns respondents to one of four cells constructed from two fully crossed factors: *treatment* (e.g., treatment and control) and *pretest administration* (present or not). Therefore, the four cells are (1) treatment with pretest, (2) control with pretest, (3) treatment without pretest, and (4) control without pretest. The analysis tests whether treatment effects are found, whether the groups who received the pretest performed better than those who did not, and whether taking a pretest interacts with treatment effects. If X (or its absence) is treatment (or its absence), and O_1 is pretest and O_2 is posttest, the design is diagrammed as follows:

Experimental Group	O_1	X	O_2
Control Group I	O_1	O_2	
Control Group II		X	O_2
Control Group III			O_2

History of the Design

Richard Solomon first proposed a three-group design that consisted of an Experimental Group that was exposed to the pretest and the treatment; Control Group I that was exposed to the pretest, but not the treatment; and Control Group II that was not exposed to the pretest but received the treatment. The three-group design served the same purpose as the four-group design, but the two designs used slightly different statistical analyses. While Solomon believed that the three-group design was sufficient for laboratory experiments, a four-group design would be better suited for field studies, including those that use sample surveys to gather data. Therefore, he added Control Group III, in which units were not exposed to either the pretest or the treatment. He originally saw the fourth group as a way to estimate effects of threats to validity other than testing, such as maturation or history, but this four-group design has become the preferred method for examining testing effects.

Analysis and Interpretation

Using this design, pretest sensitization is determined by conducting a two-factor analysis of variance (ANOVA), in which treatment and presence of pretest are the factors and posttest scores are the dependent variable. It is not necessary to include the pretest scores in the analysis because random assignment ensures that pretest scores should be the same for each group on expectation. A significant interaction between treatment and pretest suggests that exposure to the pretest may influence the treatment effect. If there is no interaction, a significant main effect for treatment indicates that the treatment had an effect, and a significant main effect for the pretest indicates that pretest may influence posttest.

Improving Validity

The Solomon four-group design provides information relevant to both internal and external validity. Regarding internal validity, for example, many single-group quasi-experiments use both pretests and posttests. If a pretest sensitization occurs, it could be mistaken for a treatment effect, so that scores could change from pretest to posttest even if the treatment had no effect at all. In randomized experiments, pretest sensitization effects do not affect internal validity because they are randomized over conditions, but those effects affect external validity if the size of the effect depends on whether a pretest is administered.

Examples

John Spence and Chris Blanchard used a Solomon four-group design to assess pretest sensitization on feeling states and self-efficacy and treatment effects when using an aerobic fitness intervention to influence feeling states and self-efficacy. The pretest and posttest questionnaires were (a) the Subjective Exercise Experience Scale, a 12-item questionnaire that measures psychological responses to stimulus properties of exercise, and (b) a self-efficacy questionnaire, which asked respondents about their ability to exercise at a high level for various time lengths. The intervention required respondents to pedal on a cycle ergometer at 60 rotations per minute for 12 minutes at three different target heart rates. There was no main effect for pretest sensitization for feeling states, nor was there an interaction between pretest administration and the treatment. However, the aerobic intervention did affect feeling states: Treated respondents reported a better sense of well-being and decreased psychological distress. There were no significant main effects or interactions for self-efficacy.

Bente Traeen used a Solomon four-group design as part of a group randomized trial, in which schools were randomly assigned to the four conditions. The intervention consisted of a sexual education course to increase contraceptive use; the pretests and posttests were self-report questionnaires asking about sexual activity. The researcher found an interaction between the pretest and intervention, indicating that the pretest may have affected the size of the treatment effect.

M. H. Clark and William R. Shadish

See also Experimental Design; External Validity; Factorial Design; Interaction Effect; Internal Validity; Main Effect; Random Assignment

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs*. Chicago: Rand McNally.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*, 137–150.

Spence, J. C., & Blanchard, C. (2001). Effect of pretesting on feeling states and self-efficacy in acute exercise. *Research Quarterly for Exercise and Sport*, *72*(3), 310–314.

Traeen, B. (2003). An effect of an intervention to prevent unwanted pregnancy in adolescents. A randomized, prospective study from Nordland County, Norway, 1999–2001. *Journal of Community and Applied Social Psychology*, *13*, 207–223.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal*, *19*, 249–258.

Spiral of Silence

Originally developed in the early 1970s by German political scientist and pollster Elisabeth Noelle-Neumann, the spiral of silence remains one of the few theoretical approaches that attempts to understand public opinion from a process-oriented perspective. The general conceptual premise of this theory is that there are different styles of communicative behavior for those who are in the majority versus those in the minority for a given issue.

According to the theory, those who are in the majority are more likely to feel confident in expressing their opinions, while those in the minority fear that expressing their views will result in social ostracism, and therefore they remain silent. These perceptions can lead to a spiraling process, in which majority viewpoints are publicly overrepresented, while minority viewpoints are increasingly withheld and therefore underrepresented. This spiral results in escalating social pressure to align with the majority viewpoint, which in turn can lead to declining expressions of minority viewpoints. As the apparent "majority" position gains strength (i.e., is expressed more and more confidently in public), those who perceive themselves as being in the minority will be more likely to withhold their opinion.

In responding to those who have attempted to explain public opinion as a rational, informed process, Noelle-Neumann has argued repeatedly that public opinion is a method of social control, akin to the latest fashion trends, which exert pressure to conform and comply. At the heart of the spiral of silence theory is the notion that people are highly motivated to fit in and get along with others. Noelle-Neumann refers to this motivation to avoid group exclusion as the "fear of isolation." Noelle-Neumann's argument for the existence of the fear of isolation is based primarily on the conformity research conducted by Solomon Asch in the 1950s. Asch found that subjects who were asked to judge the length of clearly different-sized lines were surprisingly willing to offer an obviously incorrect answer when other subjects—all of whom were confederates instructed by the experimenter to offer an incorrect judgment—had also offered incorrect judgments.

Noelle-Neumann, citing the influence of Alexis de Tocqueville, posits that most people would rather go along with an incorrect majority than strike out on their own and risk social isolation from the majority group. Exceptions are noted, however, in the form of the "hardcore," who seem to pay little mind to social isolation and are willing to hold onto and express unpopular opinions. These "avant-garde" are noted for their willingness to express viewpoints that are not popular, yet their opinions are often predictive of future trends that will later be accepted as the majority view.

In noting that individuals are constantly scanning their environments for clues regarding what others think, the spiral of silence accounts for the iterative and dynamic nature of public opinion. Through interpersonal interaction and exposure to media, people are able to gauge public opinion (referred to as a "quasi-statistical sense" by Noelle-Neumann) and use that information as a means of predicting how others will react if they decide to express or withhold certain opinions. The media comprise the second key theoretical element of the spiral of silence theory. Noelle-Neumann posited that the media are consonant and ubiquitous in their presentation of issues, resulting in a media-driven opinion context. By linking this media force with interpersonal interactions, the theory incorporates two key elements of communication research: media effects and interpersonal/group dynamics.

Measurement

Early research on the spiral of silence was almost exclusively limited to survey and interview contexts. In most survey-based studies, the key dependent variable is generally conceptualized as the respondent's willingness to express his or her opinion. Yet the operationalization of this expression has varied widely. Noelle-Neumann's original research focused on West Germans' views of East Germany. Specifically, her research sought to determine whether West Germans were for or against recognizing East Germany as a second German state, what their impressions were of the views of other West Germans, and what their perceptions were of how other West Germans would view this issue in one year's time. As Noelle-Neumann's research program expanded, survey questions were created with the goal of simulating a typical public situation in which an individual would have the choice to express or withhold his or her own opinion on a variety of political and social issues. The public situations included riding in a train compartment with a stranger and conversing at a social gathering where the respondent did not know all of the guests. By varying the opinion of the hypothetical conversation partner or group, Noelle-Neumann and colleagues were able to gauge the effect of respondents' perceptions that they were either in the minority or the majority on a given issue. Results from this initial wave of research were mixed, with some combinations of issue and setting producing a silencing effect while others produced no such effect.

In addition to replicating Noelle-Neumann's research by asking survey respondents about hypothetical opinion expression situations involving strangers, scholars have expanded on this line of work by gauging respondents' willingness to speak out by measuring voting intention, evaluating willingness to discuss issues with family and friends, assessing financial contributions to candidates, and determining respondents' willingness to express their views on camera for news broadcasts. Several of these methods have been criticized for various methodological weaknesses. For example, some scholars have noted that voting is a private act that differs significantly in its social context from expressing an opinion to another person or in a group. Expressing an opinion to family and friends introduces a variety of social-psychological considerations, including interpersonal dynamics, as well as knowledge about the opinions of other group members, as knowledge about the opinions of close friends and family members is likely to be much greater than knowledge about the opinions of other groups. Whereas financial contributions can be associated with individuals through public finance records, writing and mailing a check is certainly a behavior distinct from publicly voicing support for a candidate. Asking respondents about a hypothetical chance to express their opinion on a news broadcast evokes a social setting in which they could face recriminations for their views, but it also lacks the immediacy of reaction in comparison to an interpersonal exchange.

To assess the opinion climate in which respondents are basing their intention to express their views, survey approaches to the spiral of silence must also include some type of contextual measure in which respondents are asked about their perceptions of the opinions of those around them or in whatever specific social environment the researcher is interested. Social scientists typically take one of two approaches in assessing the climate of opinion: (1) asking respondents whether they feel most people are in agreement with them or are opposed to their personal view, or (2) asking respondents to specifically estimate what percentage or proportion of others (voters, people in your neighborhood, etc.) are in favor of or opposed to this issue or candidate. With this information, dichotomous comparisons are possible between those who feel they are in the majority versus those who feel they are in the minority on a given issue. When respondents are asked to offer specific percentages in estimating perceptions of opinion climates, researchers are also afforded the opportunity to correlate the size of majorities and minorities with respondents' willingness to express an opinion (especially if opinion expression is gauged on a continuum rather than as a dichotomous variable). Additionally, a willingness to express one's opinion also can be used in conjunction with extant opinion polls for a given issue. This approach avoids perceptual errors in individual perceptions of the opinion climate, such as the *looking glass perception* and *pluralistic ignorance*.

Inconsistent Results

In the late 1990s, Carroll Glynn and colleagues conducted a meta-analysis of spiral of silence research. Results indicated that there was a statistically significant relationship between individual perceptions that others hold similar opinions and individual willingness to express an opinion, but the relationship was quite weak across 9,500 respondents from 17 published and unpublished studies. Suggesting that hypothetical opinion expression situations might not evoke the psychological response to contextual pressures, this study called for experimentally created opinion environments in future spiral of silence work. With experimental research, respondents could be presented with the opportunity to express an opinion or to keep silent in the face of experimentally controlled majorities and minorities through the use of confederates or deliberately constructed groups. The meta-analysis also tested the impact of moderators on willingness to express an opinion; no significant relationships were found among the moderating factors that were tested.

Though no significant moderators were found in the meta-analytic review, other research has found that key variables can impact the relationship between the perceived opinion climate and respondents' willingness to express their opinions. Demographic variables such as gender, age, education, and income have been found to account for significant differences in willingness to speak out, even when controlling for the climate of opinion. Other possible moderators include political interest and knowledge, self-efficacy, the extent to which an opinion is based on moral considerations, and perceived personal "correctness" with respect to one's opinion. Contextual moderators include characteristics of the reference group in question (i.e., whose opinions constitute the opinion climate?) as well as specific characteristics about the setting in which an opinion can be expressed or withheld.

Future Applications

The spiral of silence has remained a viable approach in social science research perhaps because it addresses the inherent multi-level nature of public opinion processes. At the macro or contextual level, whether media outlets tend toward consonance or present a multi-faceted representation of events, these outlets create a mediated context in which opinions are formed and shaped. It is also within these mediated contexts that the interpersonal exchange of opinions occurs. In other words, media outlets create the backdrop for interpersonal exchanges of opinion, and these opinion exchanges can further influence the process of opinion formation—spiraling or otherwise.

With this in mind, studying the spiral of silence within a given context requires careful planning in terms of the following:

- The unit of analysis—How will opinion expression be operationalized?
- Contextual parameters—Within what context will opinion expression occur?
- Contextual characteristics—How will the opinion climate be measured?
- Individual factors—What can moderate this relationship?
- Issue contexts—What types of issues are most susceptible to the spiral of silence?

With the increasing use of multi-level modeling in the social sciences, scholars are afforded the opportunity to capture both contextual-level and individual-level variables as they impact opinion expression. After determining the contextual parameters of the opinion environment in question (e.g., small group, metropolitan area, state, or nation), it should then be possible to establish what media channels and sources are likely to produce opinion-influencing content (e.g., newspapers, television news or ads, Web sites, blogs). By content-analyzing these outlets, researchers are able to measure exactly how consonant media outlets are in presenting issue stances. This information can then be used as a contextual- or group-level variable in predicting opinion expression within a survey or experimental environment. The hierarchical structure of multi-level modeling also allows the inclusion of possible moderators that may impact the relationship between the context and opinion expression.

Recent research concerned with the spiral of silence process has focused on cultural characteristics as moderating variables in terms of contextual impacts upon willingness to express an opinion. Results have been less than consistent, but some findings indicate that collectivist cultures, which focus on face-saving techniques in the name of group harmony, are more susceptible to pressure to conform to a majority opinion (or withhold a minority opinion). This group-level result contrasts with individualistic cultures, which tend to favor personal ideas over the views of others.

Another underdeveloped area of spiral of silence research is investigation of the concept of fear of isolation. Noelle-Neumann's theory has been repeatedly criticized for assuming that all individuals—with the exception of the avant-garde and hardcore—are equally averse to social isolation and therefore will react to majority pressure in a similar way. If certain individuals are willing to express an opinion in the face of a hostile majority (e.g., the hardcore), it stands to reason that fear of isolation is a continuous concept rather than a dichotomous one.

Though some spiral of silence research has employed lagged design or panel studies in which changes in opinion can be documented, time is yet another important variable that is often overlooked when studying opinion contexts. If a true spiraling effect is to be found, opinions must be measured at multiple time points in order to gauge the effect of the pressure exerted by the climate of opinion.

Michael Huge and Carroll J. Glynn

See also Opinion Norms; Panel Survey; Public Opinion; Social Isolation

Further Readings

Glynn, C. J., Hayes, A. F., & Shanahan, J. (1997). Spiral of silence: A meta-analysis. *Public Opinion Quarterly*, *61*, 452–463.

Noelle-Neumann, E. (1993). *The spiral of silence*. Chicago: University of Chicago Press.

Price, C., & Allen, S. (1990). Opinion spirals, silence and otherwise: Applying small-group research to public opinion phenomena. *Communication Research*, *17*, 369–392.

Salmon, C. T., & Kline, F. G. (1985). The spiral of silence ten years later: An examination and evaluation. In K. R. Sanders, L. L. Kaid, & D. Nimmo (Eds.), *Political communication yearbook 1984* (pp. 3–30). Carbondale: Southern Illinois University Press.

Scheufele, D. A., & Moy, P. (2000). Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International Journal of Public Opinion Research*, *12*(1), 3–28.

Split-Half

Split-half designs are commonly used in survey research to experimentally determine the difference between two variations of survey protocol characteristics, such as the data collection mode, the survey recruitment protocol, or the survey instrument. Other common names for such experiments are *split-sample*, *split-ballot*, or *randomized experiments*. Researchers using split-half experiments are usually interested in determining the difference on outcomes such as survey statistics or other evaluative characteristics between the two groups.

In this type of experimental design, the sample is randomly divided into two halves, and each half receives a different treatment. Random assignment of sample members to the different treatments is crucial to ensure the internal validity of the experiment by guaranteeing that, on average, any observed differences between the two groups can be attributed to treatment effects rather than to differences in subsample composition. Split-half experiments have been successfully used in various survey settings to study measurement error bias as well as differences in survey nonresponse rates.

This experimental design has been used by questionnaire designers to examine the effects of questionnaire characteristics on answers provided by survey respondents. Current knowledge of question order effects, open- versus closed-response options, scale effects, response order effects, and inclusion versus exclusion of response options such as "Don't Know" is based on split-half experiments. These experiments have been conducted both in field surveys and in laboratory settings. Researchers usually assume that the experimental treatment in questionnaire split-half designs that produces the better result induces less measurement bias in the survey statistic of interest. However, these researchers often conduct such experiments because they do not have a gold standard or true value against which to compare the results of the study. Thus, the difference in the statistic of interest between the two experimental groups is an indicator of the difference in measurement error bias. Without a gold standard, it does not, however, indicate the amount of measurement error bias that remains in the statistic of interest.

Split-half experiments have also proven useful for studying survey nonresponse. Experimenters interested in how variations in field recruitment procedures—such as interviewer training techniques, amounts or types of incentive, advance and refusal letter characteristics, survey topic, and the use of new technologies such as computer-assisted self-interviewing—affect unit response, contact, and refusal rates have used split-half designs. These experiments often state that the design feature that had the higher response rate is the better outcome. Nonresponse bias is less frequently evaluated using a split-half design. Such designs have also been used to study item nonresponse as a function of both questionnaire characteristics and survey recruitment protocol characteristics.

Split-half experiments are an important experimental design when examining the effect of different survey protocol features on survey statistics. However, they do not automatically reveal which protocol or instrument choice is better. In general, to determine the "better" treatment, the researcher fielding a split-half experiment should use theory to predict the desirable outcome. For example, the type of advance letter that produces a higher response rate will often be considered superior, under the hypothesis that higher response rates lead to more representative samples. Alternatively, the mode of data collection that increases the number of reports of sensitive behaviors is considered better under the assumption that respondents underreport sensitive behaviors. Although split-half experiments are a powerful experimental design that isolate, on average, the effects that different treatments, survey protocols, or other procedures have on various types of outcomes, they are made practically useful when the survey researcher has a theory on which outcome should be preferred.

A final point of interest, despite the name split "half," survey researchers are not obligated from a statistical standpoint or methodological standpoint to actually split their samples exactly in half (i.e., 50/50). For example, a news organization may be conducting a survey on attitudes towards the Iraq war and want to determine whether alternative wording for the primary support/opposition attitude question will elicit different levels of support/opposition to the president's policies. However, the news organization may not want to "dilute" their measurement of this key attitude, for news purposes, by assigning half of their sample to the alternative wording and may choose to assign far less than 50% of respondents to the alternative wording.

Sonja Ziniel

See also Experimental Design; Internal Validity; Measurement Error; Missing Data; Nonresponse; Nonresponse Bias; True Value

Further Readings

Fowler, F. J. (2004). The case for more split-sample experiments in developing survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 173–188). New York: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Standard Definitions

Broadly, the term *standard definitions* refers to the generally accepted nomenclature, procedures, and formulas that enable survey researchers to calculate outcome rates for certain kinds of sample surveys and censuses. Specifically, *Standard Definitions* is the shorthand name for a booklet published by the American Association for Public Opinion

Research (AAPOR) titled *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Much of this entry is gleaned from that booklet.

At least two major survey organizations have formal definitions for outcome rates: AAPOR and the Council of American Survey Research Organizations (CASRO). Both organizations' formal definitions are easily accessible on the Internet. However, AAPOR provides an Excel spreadsheet response-rate calculator that makes it easy to calculate and track different outcome rates for different surveys and quickly compare rates across surveys.

Response rates—more properly known as *case outcome rates*—are survey research measures that indicate the proportion of a sample or census that do or do not responded to a survey. They are one of a number of indicators that can point to the quality of the survey. Generally, outcome rates can be broken down into four categories: cooperation rates, refusal rates, incidence rates, and overall response rates. Their calculation is based on the classification of the final disposition of attempts to reach each case in a sample, such as a household or an individual respondent.

Importance of Outcome Rates

Outcome rates of sample surveys and censuses are important to understand because they are indicators of potential nonresponse effects. Nonresponse and its effects are one of several types of potential nonrandom errors in surveys (others include coverage error and measurement error). Nonresponse bias is the extent to which the representativeness of a sample is compromised if there are nonnegligible differences between nonresponders (i.e., those who were in the originally drawn sample but who did not complete an interview) and respondents (those who did complete an interview).

A sample is a set of elements, or cases—in social science research, they usually are people or households—drawn from some population. The population may be adults in the United States, likely voters in an election, newspaper readers in a metropolitan area, college students on a campus, or some other identifiable group. The researcher measures the characteristics of these cases, usually with an instrument called a questionnaire, and in the subsequent statistical analysis infers the statistics from the sample to the population. Error bias in the sample can hamper the ability of the researcher to infer accurately the sample statistics to the population.

Just as doctors can get indicators of a patient's health by checking key components of blood chemistry, researchers can get hints at sample validity by carefully performing certain checks on the sample. One, for example, is comparing sample statistics to known population parameters: the sample's percentage of men and women, say, to U.S. Census percentages. A second is to compare the sample's findings to other, similar samples done with similar populations at about the same time. Outcome rates also can be indicators of sample validity, but they should never be used as the sole judge of a sample's quality.

Categorization of Outcomes

Just as doctors have to prepare equipment (and even the patient) before testing blood, the researcher also has to carefully categorize the outcome of each attempt to reach respondents in the sample and use that information to determine final outcome codes for each case. All this has to be done before the researcher can calculate final outcome rates. Classifying final dispositions can come at any time during the data collection period for some clear-cut cases: completed interviews, refusals, or, in random-digit dialing (RDD) samples, telephone numbers that are not working or have been disconnected.

However, during the data collection period, many attempts to reach respondents must be assigned temporary outcome codes for cases where the final disposition may not be known until the field work is over. An example for mail surveys is when the first attempt at mailing a respondent a questionnaire results in a returned envelope with no questionnaire, in which case the researcher could mail a second questionnaire with a letter asking the respondent to complete it. An example for a phone survey is one in which a respondent agreed to complete the interview but asked the interviewer to call back at a more convenient time. In both examples, the researcher likely would be using temporary outcome codes (i.e., part of what has come to be considered *paradata*) that would be useful in keeping track of the respondent sampled and would be useful in guiding the researcher later when it came time to assign final outcome codes.

Generally, final outcome categories for survey cases can be divided into (1) interviews, (2) cases that were eligible to be included but were not (nonrespondents), (3) cases that were not really eligible to be in the sample and therefore excluded from the outcome rate classifications, and (4) cases of unknown eligibility. Examples of the first two are relatively easy to understand: completed and partial interviews; and noncontacts, refusals, and language or other barriers to participation. In the third case, especially for RDD telephone samples of the general population, some of the elements clearly do not belong in the sample and are discarded: business or government telephone numbers in a sample of adults in residential households, for example.

Calculation of Outcome Rates

In many samples, it is unclear whether particular cases (e.g., household addresses, specific people, or phone numbers) really should be a part of the sample. In RDD phone samples researchers may never resolve the disposition of many telephone numbers originally included in the designated sample to which many call attempts were made but no one ever answered the phone, or those for which the researcher continually reached answering machines on which the message was unclear about the classification of the phone numbers, that is, whether they were households, businesses, or government offices.

A system of measurement needed to be developed to help researchers classify the outcomes of their attempts to reach the phone numbers, or people in mail surveys, or voters in exit polls so that they can calculate their response rates and understand if they may have issues with nonresponse bias. One of the early attempts to do this was in the 1980s by a team of researchers who developed a rudimentary way of calculating the overall response rate for CASRO. That approach was deployed by CASRO and the Market Research Association for many years. However, it provided little diagnostic information for researchers to understand the quality of their samples. One reason was because it did not allow researchers to understand the components of overall response rate, such as cooperation or contact rates. The other was because there were few standard definitions of final outcomes classifications for researchers to use when they input their outcomes into the calculator.

In the 1990s, a team of researchers working under the auspices of the American Association for Public Opinion Research, and led by Tom W. Smith, Paul J. Lavrakas, and Rob P. Daves, developed standard definitions of outcomes and used those definitions to develop a number of outcome rate formulas for RDD telephone surveys, mail surveys of individual persons, and in-home face-to-face surveys of individual persons.

The AAPOR method of calculating outcome rates allowed researchers for the first time to use standard definitions for outcomes of interviewing attempts. It also, for the first time, allowed researchers to examine the major components of nonresponse: incidence rates, cooperation rates, refusal rates, and proportions of unidentified cases. Additionally, it allowed researchers to compare response rates for one mode, such as an RDD phone survey, to another mode, such as a mail survey.

Classification Problems

One of the problems that plague survey researchers is when case outcomes are not clear-cut. For example, in face-to-face interviews with people in their own households, a researcher may never find any eligible respondent at home during the field period. There may clearly be an eligible respondent, but the interviewer is never able to talk to him or her because he or she is unavailable during the data collection period. This is a case of nonresponse, but it cannot be classified as a respondent's refusal.

Some telephone survey case outcomes are even harder to classify. If the researcher uses an RDD sample of landline telephone numbers, many will be business, government, disconnected, fax or computer lines, or not working; these clearly are not considered to be part of the eligible telephone numbers for a residential sample and are excluded from outcome rate calculations. But what happens, for example, when during the data collection period, interviewers call a dozen times and it always rings but no one answers? This example falls into a group of outcomes that force the researcher to make assumptions about the proportion that might be truly eligible. There are many ways to deal with this. One is to assume that the group of phone numbers with an unknown eligibility split in the same proportion of eligible-to-ineligible phone numbers in the group with known eligibility in the sample. This may not be always accurate, but it may be the best that the researcher can do. The AAPOR standard definitions and formulas take this into account. Certain AAPOR formulas use this assumption as the default.

Outcome Rate Categories

Outcome rates generally fall into four categories. They are as follows:

1. *Contact rates*—the proportion of all cases in the sample (or census) in which some responsible housing unit member was reached
2. *Refusal rates*—the proportion of all cases in which there is a refusal to be interviewed
3. *Cooperation rates*—the proportion of all eligible units reached in which there was a completed interview
4. *Response rates*—The number of complete interviews divided by the number of eligible units in the sample

In 2001, AAPOR researchers put the *Standard Definitions* guide on AAPOR's Web site; it was updated in 2006. They also developed an Excel spreadsheet that allows other researchers to quickly and accurately enter their final case dispositions and calculate the various outcome rates. (This spreadsheet is downloadable free at the AAPOR Web site.)

Sample Validity

Because using standard definitions of case outcomes is relatively recent and their adoption has not been universal, AAPOR outcome rates appear to be used more among academic researchers and those who do media polls and less among commercial market researchers. Researchers have only just begun exploring how differences in outcome rates affect sample validity. Some early research suggests that extremely high outcome rates in samples of likely voters actually can hinder the ability of public opinion researchers to be as accurate as they need to be in pre-election polls. Other research for general population samples suggests that validity of samples for some types of measures response rates between 40% and 70% does not hamper validity. In fact, a team at the University of Michigan, the University of Maryland, and the Pew Center for the People and the Press found that for many measures of political attitudes, there were few differences between a survey conducted with a short data collection period (and thus a low response rate) and one in which the data collection period was greatly extended with the use of many extraordinary techniques to increase response rates. However, it is crucial to understand that researchers are continuing to investigate the effect of nonresponse on sample surveys, and as yet there is no conclusive evidence.

Robert P. Daves

See also American Association for Public Opinion Research (AAPOR); Council of American Survey Research Organizations (CASRO); Dispositions; Elements; Nonresponse; Paradata; Response Rates; Sample

Further Readings

American Association for Public Opinion Research: http://www.aapor.org

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Ann Arbor, MI: Author. Retrieved April 19, 2008, from http://www.aapor.org/uploads/standarddefs_4.pdf

Frankel, L. R. (1982). *On the definition of response rates: A special report of the CASRO Task Force on Response Rates*. Retrieved April 19, 2008, from http://www.casro.org/resprates.cfm

STANDARD ERROR

Statistics are derived from sample data, and because they are not taken from complete data, they inevitably vary from one sample to another. The standard error is a measure of the expected dispersion of sample estimates around the true population parameter. It is used to gauge the accuracy of a sample estimate: A larger standard error suggests less confidence in the sample statistic as an accurate measure of the population characteristic. Standard errors can be calculated for a range of survey statistics including means, percentages, totals, and differences in percentages. The discussion that follows focuses on sample means and proportions.

In a survey of families, let X represent family income, and let $\bar{X}$ denote the mean family income, which is an example of a statistic resulting from the data. Although the survey may be designed to provide an estimate of the mean in the entire population of interest, it is highly unlikely that the estimate from the survey, $\bar{X}$, will be equal to μ, the mean income of all families in the targeted population.

Assuming that each family in the population has an equal chance of being selected for the survey, the true standard error of $\bar{X}$ could be derived, hypothetically, by conducting an infinite number of identical surveys of independent samples of the same size from the same population. The distribution of the values of $\bar{X}$ comprises the sampling distribution of $\bar{X}$. The mean of this sampling distribution is the true value of the parameter, that is, μ, that the statistic is meant to estimate, and the standard deviation of the sampling distribution is the true standard error associated with the statistic. It can be denoted by $\sigma_{\bar{x}}$.

In practice, however, the true population mean μ and the true standard error $\sigma_{\bar{x}}$ of $\bar{X}$ are not derived from repeated surveys. In fact, they are rarely known. Instead, they are estimated from a single sample, and it is this estimate of the standard error that has become synonymous with the term *standard error* in survey research. With this understanding, the standard error of $\bar{X}$, denoted $s_{\bar{x}}$, is defined in the following section.

The Standard Error of a Sample Mean

If s is the standard deviation of the family income variable X, and n is the number of households surveyed, the standard error of $\bar{X}$, the mean family income, is estimated as $s_{\bar{x}} = s/\sqrt{n}$. For example, if the sample size is 100, and the standard deviation of family income is $15,280, then the standard error is equal to $1,528 because $15,280/\sqrt{100} = 15,280/10 = 1,528$.

The standard error, by itself, is not easily interpreted, and that is why confidence intervals and margins of error are more often reported with a survey statistic. These measures are closely related to the standard error, but, perhaps, offer a more intuitive interpretation of the uncertainty of a survey result. Recall the true standard error of $\bar{X}$ is the standard deviation of sample means around the true mean in the population. When the sample size is large, the sampling distribution of $\bar{X}$ is normally distributed, which means approximately 68% of all values will lie within one standard deviation of the true mean μ. Therefore 68% of all sample estimates of μ will lie within the interval $\bar{X} +/- s_{\bar{x}}$, which is the usual estimate for the 68% confidence interval for μ. If the mean family income resulting from the survey is $\bar{X} = 37,500$, and its standard error is $s_{\bar{x}} = 1,528$, then the 68% confidence interval for μ is $37,500 +/− $1,528 or the interval, $35,972 − $39,028.

Given $\bar{X}$ and $s_{\bar{x}}$, any confidence interval can be calculated, but it is conventional in survey research to report the 95% confidence interval. If 68% of the values in a normal distribution fall within one standard deviation of the mean, then 95% of the values lie within 1.96 standard deviations of the mean. The 95% confidence interval is expressed, then, as $\bar{X} +/- 1.96 * s_{\bar{x}}$. The margin of error is usually defined as the radius of the 95% confidence interval, or $1.96 * s_{\bar{x}}$. In the example of family income, where $\bar{X} =$ $37,500 and $s_{\bar{x}} =$ $1,528, the 95% confidence interval is $37,500 +/− $2,994.88 and the margin of error is $2,994.88.

The Standard Error of a Sample Proportion

In surveys and polls the result of interest is often the proportion of sample elements that belongs to a class or possesses an attribute, denoted $\hat{p}$. For example, in a survey of families, it might be important to estimate the proportion of families with two parents in residence. If $\hat{p}$ is the sample proportion possessing the attribute, then $\hat{q}$ is the proportion not possessing the attribute. The standard error of $\hat{p}$ is defined as $s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$. If the survey found that 80 of the 100 families had two parents present, then $\hat{p} = 80/100 = 4/5$ and $\hat{q} = 20/100 = 1/5$. The standard error of $\hat{p}$ would be

$$s_{\hat{p}} = \sqrt{\frac{\frac{4}{5} * \frac{1}{5}}{100}} = \sqrt{\frac{\frac{4}{25}}{100}} = .04.$$

Confidence intervals could be constructed for p in the same way as for μ (shown in the previous section). The 95% confidence interval for p is estimated as $\hat{p} +/- 1.96 * s_{\hat{p}}$ and the margin of error is $1.96 * s_{\hat{p}}$. The 95% confidence interval for the proportion of families with two parents is then $.8 \ +/- 1.96 * .04$, which is $.8 +/- .0784$. The margin of error is .0784.

The Standard Error and Sample Size

In the calculations presented previously, the standard errors for the sample means and proportions would decrease in magnitude if sample sizes were larger; that is, the n in each formula is a denominator, and as it grows larger, the standard error grows smaller. In general, larger samples have less random sampling error and provide more accurate estimates of the unknown population parameter. Sample size also affects the accuracy of confidence intervals and margins of error estimates. These are usually approximated by assuming that the distribution of sample estimates is a normal distribution. This assumption is warranted for the distributions of $\bar{X}$ and $\hat{p}$ when the sample size is large—generally greater than 30. However, when the sample size is small, the normality assumption is a stronger assumption, and estimates of confidence intervals and margins of error based on the assumption of normality are more questionable.

The Standard Error, Sampling Strategy, and Survey Methods

Sampling design and survey methods influence the accuracy of sample estimates and therefore affect the magnitude of the standard error. The methods described here for calculating standard errors assume a simple random sample, in which every element in the targeted population has an equal probability of being selected. When the sample is not a simple random sample but rather a more complicated probability sample, standard errors and confidence intervals must be estimated through more advanced techniques. In any case, standard errors are measures of sampling error. They do not take into account other sources of error such as a nonrepresentative sampling, poorly phrased questions, untruthful responses, missing data due to "don't know" or "undecided" responses, and overall response rates.

Jani S. Little

See also Confidence Interval; Equal Probability of Selection; Margin of Error (MOE); Population Parameter; Probability Sample; Sampling Error; Simple Random Sample; Standard Error of the Mean; Statistic; Variance

Further Readings

American Statistical Association, Survey Research Methods Section: http://www.amstat.org/sections/srms/

Devore, J. L. (1991). *Probability and statistics for engineering and the sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kish, L. (1995). *Survey sampling* (Reprinted ed.). New York: Wiley. (Original work published 1965)

Section on Survey Research Methods. (1998). *What is a survey? What is a margin of error?* Alexandria, VA: American Statistical Association. Retrieved on May 29, 2008, from http://www.whatisasurvey.info

Standard Error of the Mean

The standard error of the mean refers to the standard deviation of the sampling distribution of the sample mean. This distribution represents all possible sample means that could be computed from samples selected according to a specified sample size and sampling design. The standard error of the mean quantifies how much variation is expected to be present in the sample means that would be computed from each and every possible sample, of a given size, taken from the population. The standard error of the mean is measured in the same units as the original data and is often denoted by $SE(\bar{X})$ or simply as SE. Larger SE values imply more variation in sample means across possible samples of the same size; smaller SE values imply that the sample mean is more precise, or varies less from one sample to another.

The SE is typically estimated by dividing the estimate of the population standard deviation by the square root of the sample size: $SE(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}}$. Generally, increases in sample size imply decreases in the SE. Additionally, the SE is usually much smaller than the sample standard deviation with the degree of the difference being inversely proportional to the square root of the sample size.

In the context of a simple random sample of size n, selected without replacement from a finite population of size N with a population standard deviation σ, the standard error of the mean is given by

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}},$$

where the last part of the formula represents the "finite population correction." If the population is much larger compared to the actual size of the sample, then the two SE formulas will be approximately equal. If σ is unknown, it can be estimated using information from the latest version of the survey (i.e., estimate from previous cycle) or from the sample (i.e., sample standard deviation).

Uses of the Standard Error of the Mean in Survey Research

The design effect for the mean for a given survey sampling design is the square of the quotient of the standard error of the mean, based on the particular design, to the standard error of the mean based on simple random sampling without replacement.

Confidence intervals for the population mean are also computed using the sample mean estimate along with an estimate of the standard error of the mean. Typically formulas for a $(1-\alpha) \times 100\%$ confidence interval for the population mean are presented in the form

$$\bar{X} \pm \textit{Critical Value} \times SE(\bar{X}),$$

where the "critical value" is computed according to some statistical reference distribution such as the standard normal or t-distribution.

The coefficient of variation is simply the quotient of the standard error of the mean to the sample mean. Because the standard error of the mean is influenced by the units of the data, the coefficient of variation allows researchers to compare variability in the sample means across different samples using the same variables that are perhaps measured on different scales, such as income ranging in the thousands of dollars compared to income measured in millions of dollars.

For example, suppose that there is an interest in estimating the mean number of days in the past year that teenagers living in a rural community consumed at least one alcoholic beverage, and data from a probability sample, such as that provided by the National Survey on Drug Use and Health, are used to make this estimate. Assuming that a simple random sample of 100 teenagers from the rural community of 1,000 teenagers produces a sample mean of 25.75 and a sample standard deviation of 30.0 days, then the estimated *SE* is $SE(\bar{X}) = 30/\sqrt{100} = 3$. Using the finite population version, the estimated *SE* becomes

$$SE(\bar{X}) = \frac{30}{\sqrt{100}} \times \sqrt{1 - \frac{100}{1000}} = 2.846.$$

Trent D. Buskirk

See also Confidence Interval; Design Effects (*deff*); Finite Population Correction (fpc) Factor

Further Readings

van Belle, G. (2002). *Statistical rules of thumb*. New York: Wiley.

Yates, D., Moore, D., & McCabe, G. (1999). *The practice of statistics*. New York: W. H. Freeman.

Standardized Survey Interviewing

In a standardized survey interview, the interview proceeds according to a script (the introduction and the questionnaire) that is intended to minimize any potential impact of individual interviewers' behavior on respondents' answers and the resulting data. Standardized interviewing procedures for sample surveys were developed over several decades of the 20th century as evidence accrued that even seemingly minor differences in how interviewers behaved in interviews sometimes affected answers and data quality. Interviewers' biases or assumptions about particular types of respondents could creep into the interview through subtle changes in wording or tone that could lead respondents to interpret questions or the situation differently than they would have with a different interviewer. Even without intending to influence answers, interviewers who attempted to increase rapport by rephrasing a question the second time they asked it, or politely did not present all the response alternatives to a question because they judged that some alternatives would not fit a particular respondent's circumstances, could harm the quality of the data. It also became clear that interviewers could, in all innocence and with good intentions, introduce bias through how they reacted when a respondent expressed reservations or uncertainty about an answer; interviewers could subtly encourage respondents to give answers that fit the interviewers' preconceptions rather than the respondent's actual circumstances or opinions.

Standardized survey interviewing procedures are designed to circumvent these problems and to ensure that the data from all respondents are fully comparable because all respondents have answered the same questions under the same procedures. Standardizing the interviewing procedures is intended to address the measurement error due to interviewers, which is assumed to be independent of measurement error due to question wording (which can be addressed through better question pretesting) and measurement error due to respondents (which cannot easily be addressed by survey researchers).

Ideally, interviewers adhering to standardized procedures read (either from paper or from a computer screen) survey questions and all response alternatives precisely as worded by the designers of the survey, and they repeat the full question and all response alternatives when asked to repeat the question. In the strictest forms of standardized survey interviewing, interviewers also leave the interpretation of questions entirely up to respondents and only respond to any requests for clarification with neutral probes like "whatever it means to you" and "let me repeat the question." The logic is that if only some respondents receive clarification or help with answering, then the stimulus (the question wording and response alternatives) is different for a different respondent, and thus there is no guarantee that the data are comparable.

The broad consensus is that standardized interviewing procedures are the most desirable for sample surveys and that more idiosyncratic or ethnographic forms of interviewing that are useful for other more qualitative research purposes are risky or undesirable in surveys. But this consensus can manifest itself somewhat differently in different survey organizations, where the precise procedures that count as "standardized" in one center can differ from those in other organizations. For example, organizations differ on whether providing clarification to a respondent counts as nonstandardized or standardized and on whether repeating only the most appropriate response alternatives is better than repeating them all. Survey organizations can also vary in how extensively they train and monitor their interviewers for adherence to the standardized procedures they advocate, which means that in practice some standardized surveys turn out to be less standardized than others.

Controversies

Starting in the 1990s, methodological researchers who closely examine interview recordings began documenting that strictly standardized procedures sometimes can create uncomfortable interactions that not only frustrate respondents but also lead to demonstrably poorer data quality. When interviewers "force" respondents into answers that do not reflect their circumstances, when they repeat information that the respondent already knows, when they ask for information the respondent has already provided, or when they refuse to clarify what their questions mean, respondents can become alienated and recalcitrant, and they can provide incorrect answers. When viewed through this lens, the practical results of strict standardization can, on occasion, run counter to the intended effects. Perhaps, in the attempt to standardize wording in the interview, survey researchers are failing to standardize what really ought to be standardized: the "meaning" of the questions. Within this view, to serve the goal of making sure that respondents' answers are truly comparable, interviewers should instead work to make sure that respondents are interpreting questions in the same way—even if this means deviating from a script and tailoring clarification and probes to individual respondents.

The jury is still out on the implications of this work and what it will mean for the future of standardized interviewing. There are a number of important considerations. First, much is unknown about how often problematic interactions occur and how often respondents' interpretations of questions differ from interpretations of the survey designers in different domains of questioning; if problems and misinterpretations are too frequent, then alternate ways of implementing standardized interviews are worth investigating, but if they are quite rare then revamping procedures on a large scale is clearly not worthwhile. Second, the survey enterprise currently relies on a particular model of hiring and training interviewers, and it is not yet clear how training and monitoring interviewers who implement a less scripted version of standardization would work. Based on some reports, some interviewers would prefer to implement less strictly standardized interviews, but whether a different kind of interviewer (e.g., those with higher levels of education or subject matter expertise) and different levels of compensation would be needed to do this is unknown.

The larger question about the best way to achieve the goals of standardization remains: Should survey researchers standardize the stimulus (question) and the subsequent interaction? Or should survey researchers standardize the respondent's interpretation of the question and experience of the interaction?

This question is put into sharp relief by comparing standardized interviewing with self-administered surveys. In a sense, a self-administered survey (whether administered via paper and pencil, via clicking on a textual Web survey, or via listening to a recorded voice on a laptop computer) is the ultimate in standardization: All respondents are presented with precisely the same stimulus, with no variation in what the "interviewer" (words on the page, audio recording) does. One could argue that the future of standardized interviewing is to migrate human interviews into automated self-administered interviews, because, in essence, this creates one interviewing agent for the entire survey sample, and so the differential impact of different interviewers disappears. (The less-quantified aspects of human contact and rapport building would also, of course, disappear.)

But the larger question about what it takes to make survey responses comparable and minimize the effects of the interviewer or interviewing agent would still not be addressed, even with full self-administration. For example, in a paper-and-pencil survey, the interpretation of the questions is left entirely up to the respondent, but should it be? Would the data be more comparable if respondents were all assisted in interpreting the questions in the same way (e.g., via clarification dialogue in a Web survey)? In an audio self-administered survey, will different respondents interpret the recorded interviewer's vocal tone differently in ways that affect their answers? Would the data be more comparable if the interviewer's voice were variably tailored so as to be experienced as more similar across respondents?

Theories and practices of standardization—and thus questions about standardization—are thus expanding rapidly with new interviewing technologies. Not only will debates about, and data from, studies of human interviews inform those new technologies, but also the careful manipulation of features of human interaction that will come from new technologies will inform ongoing debates about standardization in human interviewing.

Michael F. Schober

See also Audio Computer-Assisted Self-Interviewing (ACASI); Conversational Interviewing; Dependent Interviewing; Interviewer Monitoring; Interviewer-Related Error; Interviewer Training; Introduction; Measurement Error; Probing; Questionnaire; Respondent–Interviewer Rapport; Tailoring; Verbatim Responses

Further Readings

Beatty, P. (1995). Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics*, *11*, 147–160.

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, *64*, 1–28.

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Newbury Park, CA: Sage.

Houtkoop-Steenstra, H. (2000). *Interaction and the standardized survey interview: The living questionnaire*. Cambridge, UK: Cambridge University Press.

Maynard, D., Houtkoop-Steenstra, H., Schaeffer, N. C., & van der Zouwen, J. (Eds.). (2002). *Standardization and tacit knowledge: Interaction and practice in the survey interview*. New York: Wiley.

Suchman, L., & Jordan, B. (1990.) Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, *85*, 232–253.

STATA

STATA is a general-purpose interactive statistical software package available in major platforms such as Windows, Unix, and Macintosh. In part due to its up-to-date coverage of statistical methodology and flexibility in implementing user-defined modules, STATA has gained considerable popularity among social and behavioral scientists, including survey researchers, in recent years despite its initial learning curve for the uninitiated.

STATA comes in four versions: (1) small STATA, a student version; (2) intercooled STATA, the "standard" version; (3) STATA/SE, a version for large data sets; and (4) STATA/MP, a parallel-processing-capable version of STATA/SE. Depending on size of data and number of variables as well as computer capacity, most survey researchers will likely choose Intercooled STATA and STATA/SE, in that order of preference.

With a fairly developed graphics capacity, STATA offers a vast array of commands for all kinds of statistical analysis, from analysis of variance to logistic regression to quantile regression to zero-inflated Poisson regression. Although not an object-oriented language like C, R, or S, STATA is fairly programmable, and that is why there is a huge collection of user-written macros, known as *ado-files*, supplementing the main program of STATA and which are typically well documented and regularly maintained. These ado-files satisfy a spectrum of needs among common users. Two examples provide a sense of the range: *SPost*, which is a set of ado-files for the post-estimation interpretation of regression models for categorical outcomes, and *svylorenz*, which is a module for computing distribution-free variance estimates for quantile group share of a total, cumulative quantile group shares (and the Gini index) when estimated from complex survey data.

STATA already has good capacity for analyzing survey data in its main program. For example, the *svyset* command declares the data to be complex survey type, specifies variables containing survey design information, and designates the default method for variance estimation. Many regression-type commands work with the cluster option, which gives cluster-correlated robust estimate of variance. Adjustment for survey design effects can also be achieved by using the *svy* prefix command (e.g., *svy*: logit) before a command for a specific operation. STATA supports three major types of weight: frequency weight (*fweight*) denoting the number of duplicated cases, probability or sampling weight (*pweight*) for indicating the inverse of the probability that the observation is included due to sampling design, and analytic weight (*aweight*) being inversely proportional to the variance of an observation. (Another weight, importance weight, can be used by a programmer for a particular computation.)

The *xt* series of commands are designed for analyzing panel (or time-series cross-sectional) data. These, coupled with the reshape command for changing the data from the wide to the long form (or vice versa) when survey data from multiple panels are combined into one file, are very attractive features of STATA for the survey data analyst.

Tim F. Liao

See also SAS; Statistical Package for the Social Sciences (SPSS)

Further Readings

Hamilton, L. C. (2006). *Statistics with STATA*. Belmont, CA: Duxbury.

Statistic

A statistic is a numerical summary of observed values for a random variable (or variables) in a sample. The term *random* indicates that the variable's value, and thus its statistic, may differ across samples that are drawn from the same sample population. A statistic is used to estimate a parameter, a numerical summary of a given variable (or variables) in the population.

A statistic is commonly represented with the common alphabet rather than with Greek letters, as is typically the case with a parameter. For example, a statistic such as a standard deviation is represented by s, whereas the corresponding population parameter is represented by σ.

Describing characteristics of a sample variable (or variables) is often called *descriptive statistics*. Examples include computing the sample's mean or standard deviation on a variable such as age or weight or describing proportions for categorical variables such as race or marital status. The process of using the sample's descriptive data to generate population estimates is typically referred to as *inferential statistics*.

It should be recognized that if data are available for the entire population of interest, inference is unnecessary as the parameter can be calculated directly.

Kirsten Barrett

See also Inference; Parameter; Population; Random; Sample

Further Readings

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice Hall.

Statistical Package for the Social Sciences (SPSS)

In 1968, Normal Nie and Dale Bent, Stanford doctoral candidates, and Hadlai Hull, a Stanford graduate with a master's in business administration, developed a software system that allowed for the transformation of raw data into information using statistical applications. Their creation, the *Statistical Package for the Social Sciences (SPSS)*, was not developed with the intent of mass distribution. However, its appeal quickly caught on and, soon after its development, it was in high demand in universities across the United States. This demand further increased after McGraw-Hill published the first SPSS user's manual in 1970. SPSS incorporated in 1975.

SPSS, Inc., offered the first mainframe statistical package to appear on a personal computer (in the mid-1980s) and, in 1992, was the first organization to release a statistical package for use with the Microsoft

Windows operating system. In the 21st century, SPSS, Inc., now offers a broad array of products and services to meet the diverse needs of its customers, both in the United States and abroad. The flagship product for SPSS, Inc., is SPSS Base, available for both Windows and Apple platforms. This product provides survey researchers with a powerful and user-friendly data management and statistical analysis package. SPSS Base allows the researcher to generate both descriptive and bivariate statistics. Further, with SPSS Base, the survey analyst can run predictive analytics such as factor and regression analyses.

A number of add-on modules and stand-alone products further enhance the capabilities of SPSS Base. Add-on modules allow for advanced multivariate analysis of survey data, including data derived from surveys with complex sample designs. Included are modules that allow for generalized linear models, hierarchical linear models, survival analysis, and categorical regression. SPSS, Inc., stand-alone products serve to add power to the data management and analysis system provided in SPSS Base. Stand-alone products are available to help with all phases of the survey process, including sample selection and data collection, data management and cleaning, data analysis, and data dissemination.

Kirsten Barrett

See also SAS; STATA; SUDAAN

Further Readings

SPSS, Inc. (2008). *About SPSS, Inc.—Corporate history*. Retrieved February 5, 2007, from http://www.spss.com/corpinfo/history.htm

Statistical Power

The probability of correctly rejecting a null hypothesis that is false is called the statistical power (or simply, power) of the test. A related quantity is the Type II error rate (β) of the test, defined as the probability of *not* rejecting a false null hypothesis. Because power is based on the assumption that the null hypothesis is actually false, the computations of statistical power are conditional probabilities based on specific alternative values of the parameter(s) being tested. As a probability, power will range from 0 to 1 with larger values being more desirable; numerically, power is equal to $1 - \beta$.

The statistical power is also related implicitly to the Type I error rate (α), or significance level, of a hypothesis test. If α is small, then it will be more difficult to reject the null hypotheses, implying that the power will also be low. Conversely, if α is larger, then the null hypotheses will have a larger rejection region, and consequently the power will be larger. While power and Type I error rates do covary as these extremes suggest, the exact relationship between power and α is more complex than might be ascertained by interpolating from these extreme cases.

Statistical power is usually computed during the design phase of a survey research study; typical values desired for such studies range from 0.70 to 0.90. Generally many survey items are to be compared across multiple strata or against some prior census value(s). For example, researchers may use data from the Current Population Survey to determine if the unemployment rate for California is lower than the national average. Power calculations can be computed for each questionnaire item, and the *maximum sample size required* to achieve a specified power level for any given question becomes the overall sample size. Generally, in practice, one or two key items of interest are identified for testing, or a statistical model relating several of the items as predictors and others as key independent variables is specified. Power calculations to determine the adequacy of target sample sizes are then derived for these specific questionnaire items or particular statistical tests of model parameters.

Consider a scenario involving a random-digit dialing sample of households selected to estimate the average food replacement costs after an extended power outage for the residents within a midwestern U.S. county. The average food loss cost per household based on data from previous storms was \$500.00 (μ_0). Because this particular storm was slightly more severe than a previous storm, officials believe that in actuality, the average food loss cost for households for the current storm is somewhere closer to \$550.00 (μ_1). The standard deviation of the distribution of food loss costs was assumed to be \$100.00. The statistical power for the one-tailed hypothesis test based on a sample of 25 houses using a Type I error rate of 5% is to be computed. In this case, the particular *effect size* used in the computation of statistical power is Effect Size $= \frac{|500 - 550|}{100} = 0.50$. The statistical power for this test is 80.51%, which represents the probability of rejecting the null hypothesis of average food loss of $500.00 given that the actual average food loss costs is $550.00 using an estimated standard deviation of 100, a sample size of 25, and $\alpha = 0.05$. Thus, there is roughly an 81% chance for detecting a positive difference in the average food loss costs of $50.00 using this hypothesis test. This power calculation is depicted graphically in Figure 1.

Notice that the Type I error rate (darker gray area) is computed with respect to the distribution defined by the null hypothesis (the curve on the left). Specifically, the Type I error rate is the area under the curve defined by the *null* hypothesis to the right of the critical value that defines the null hypothesis rejection region. On the other hand, the statistical power (light gray region) is computed as the area to the right of the null hypothesis rejection region under the curve that is defined by parameters that are given in the *alternative* hypothesis (gray/dashed curve line).

No matter what the form of the comparison (i.e., one/two sample or advanced statistical model), the power of the test is generally a function of three key values:

1. Level of significance/Type I error rate of the particular statistical test
2. The effect size—defined as the standardized difference between the null and alternative values for the parameter (which assumes some knowledge or estimate of the population standard deviation)
3. The sample size

Several Web applets are available for computing power for simple designs (e.g., http://davidmlane.com/hyperstat/power.html). Typically, these computations ignore the finite population correction. In addition to computing power values for a specific sample or effect size, statistical calculators can generate "power curves" that graphically plot the power of a statistical test (y-axis) as a function of either the effect or sample size (x-axis), for a fixed significance level. An example of a series of power curves

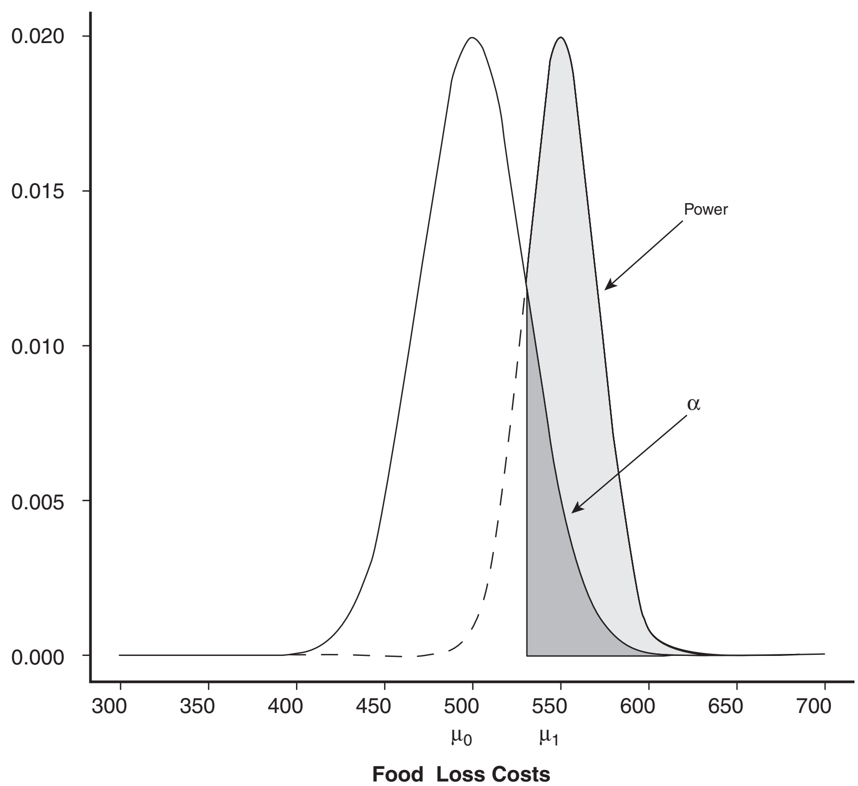

Figure 1 Illustration of the statistical power computation for a one-tailed hypothesis test for a single population mean assuming the Type I error rate is .05

computed for three different effect sizes for one-sided hypothesis tests involving a single mean is depicted in Figure 2. Notice from the figure that the statistical power increases as the sample size increases. Also notice that for any given possible sample size, the power increases as the effect size increases. Finally, note that the height of the solid curve in Figure 2 (corresponding to an effect size of .50) at the possible sample size of 25 is approximately .81, consistent with previous calculations.

Statistical power computations are to hypothesis testing of parameters as precision is to interval estimation of parameters. For example, if the goal of a survey sample is to produce reliable statewide estimates of obesity, average consumer spending per month, or other variables of interest, then precision is the defining computation that drives estimates of sample size. However, if interest is given to making comparisons for average obesity rates across several strata or post-strata, or in comparing average unemployment rates from one year to another within a state, then statistical power computations can be useful in understanding how likely one is to detect a given effect with a particular sample size. Of course, these quantities are closely intertwined statistically, but typically they imply different uses of the data at hand.

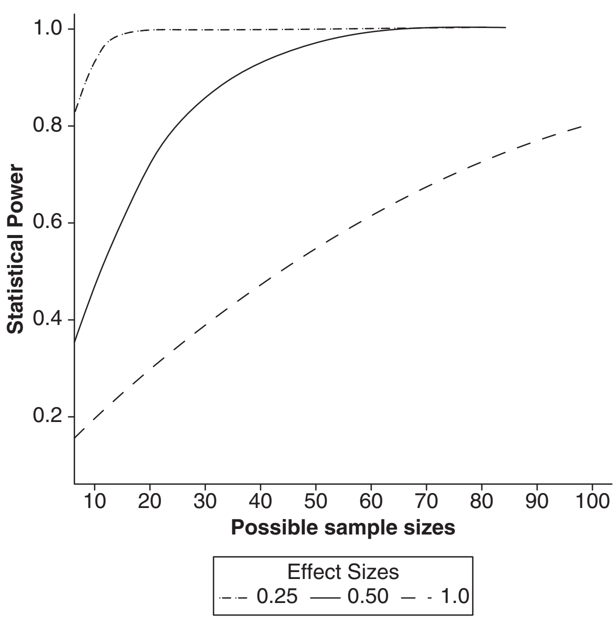

Figure 2 Statistical power curves illustrating how statistical power can be influenced by both possible sample size as well as effect size assuming a fixed Type I error rate of .05

Trent D. Buskirk

See also Alternative Hypothesis; Finite Population Correction (fpc) Factor; Null Hypothesis; Sample Size; Type I Error; Type II Error

Further Readings

D'Agostino, R. B., Sullivan, L. M., & Beiser, A. S. (2006). *Introductory applied biostatistics*. Belmont, CA: Thomson.

Statistics Canada

Statistics Canada is Canada's national statistical agency. Prior to 1971, Statistics Canada was known as the Dominion Bureau of Statistics. The bureau was created in 1918 as a permanent home for the national census and to develop a standardized form and database for the collection of vital statistics. In addition to the census and collection of vital statistics, Statistics Canada administers more than 400 surveys and compiles a number of different administrative data sources. Some examples of their cross-sectional surveys include the Labour Force Survey, Survey of Household Spending, and the General Social Survey. Statistics Canada also completes several longitudinal surveys, such as National Population Health Survey—Household Component, National Graduates Survey, National Longitudinal Survey of Children and Youth, and the Survey of Labour and Income Dynamics. These surveys about employment, income, education, health, and social conditions are a handful of the many Statistics Canada surveys.

Many census data reports are free to the public through Statistics Canada's online data source Community Profiles. CANSIM, another online data source, includes summary tables from many of Statistics Canada's surveys and administrative data sources. Some of these summary tables are free, and other tables have a nominal fee. Government officials, academics, the media, and members of the general public receive daily updates about these data sources through *The Daily*, Statistics Canada's official release bulletin.

Census

The first national census was in 1871, a few years after Confederation (1867). The census included the four original provinces and expanded as additional provinces joined Confederation. When the prairie provinces joined Confederation, an Agricultural Census was initiated (1906) and repeated every 5 years to help monitor the growth of the west. Since 1956, these two censuses have been administered together every 5 years. Agricultural producers across the country received a copy of the Census of Population questionnaire as well as a copy of the Census of Agriculture questionnaire. Participation in the census is required by law.

The intent of the first national census was to determine population estimates to ensure representation by population in Parliament, a purpose still relevant today. In Canada, the census is also used to calculate transfer payments between the different government jurisdictions (federal government to the provinces and territories; provinces and territories to municipalities). The transfer payments fund schools, hospitals, and other public services.

Prior to the 1950s, the census involved multiple questionnaires and ranged from 200 to 500 questions. To reduce response burden and administration costs, sample surveys were initiated to gather certain data and the census questionnaires were reduced in size. Two versions of the census questionnaire were introduced. The short version is administered to all Canadian households, except every fifth household, which receives a long version of the questionnaire. The short version gathers information about names of household members, relationship to the head of household, sex, age, marital status, and language. The long version includes these questions as well as additional questions about education, ethnicity, mobility, income, and employment. The content of the long version is more apt to change across time, with the short version staying more consistent in content.

In 1971, several innovations were introduced to the census. First, one in three households received the long questionnaire, but this innovation was abandoned in subsequent years in favor of the one in five sampling approach. Second, the census moved from being interviewer-administered to being self-administered for most of the population. The latter innovation has continued, and in 2006, this self-administered mode included the option of completing the questionnaire online. In 2006, almost 20% of Canadian households opted to complete the survey online. The Census 2006 also contained an additional innovation: As part of ensuring more ethical standards in data collection, the Census 2006 included a question requesting permission to release census responses to the public in 92 years.

Shelley Boulianne

See also Census; *Survey Methodology*

Further Readings

Statistics Canada. (1993). *75 Years and counting: A history of Statistics Canada*. Ottawa, ON: Minister Responsible for Statistics Canada. (Statistics Canada Catalogue no. 11-531).

Statistics Canada. (2007, March). *History of the census of Canada*. Retrieved May 28, 2007, from http://www.statcan.ca/english/census96/history.htm

Step-Ladder Question

A step-ladder question refers to a type of question sequence that yields more complete and accurate data than would a single question on the same topic. Step-ladder questions are used by survey researchers in an attempt to reduce item nonresponse (missing data), measurement error, or both, although they add slightly to the cost of gathering the data since more than one question needs to be asked.

For example, asking someone into which of the following income categories his or her 2007 total household income fell—less than $20,000; $20,000–$39,999; $40,000–$59,999; $60,000–$79,999; $80,000–$99,999; $100,000 or more—will lead to a good deal of "Don't Know" or "Refused" answers. Researchers have found that a step-ladder question about income will substantially reduce item nonresponse and thus the need to impute those missing values.

A step-ladder question sequence for the income variable referenced in the previous paragraph, that was programmed to be asked in a computer-assisted interview, would be as follows:

Q1. *Was your total household income from all sources in 2007 more than $19,999?*

< 1 > YES (GO TO Q2)
< 2 > NO (GO TO Q6)

< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)
Q2. *And was it more than $39,999?*
< 1 > YES (GO TO Q3)
< 2 > NO (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)
Q3. *And was it more than $59,999?*
< 1 > YES (GO TO Q4)
< 2 > NO (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)
Q4. *And was it more than $79,999?*
< 1 > YES (GO TO Q5)
< 2 > NO (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)
Q5. *And was it more than $99,999?*
< 1 > YES (GO TO Q6)
< 2 > NO (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)

In this example, after the income sequence has been administered all respondents are taken to Q6 (i.e., whatever is the next logical question in the questionnaire after the income sequence). Of note, even though the entire step-ladder sequence comprises five questions, any one respondent would be asked more questions in the sequence only up until he or she said "No." As such, proportionally few respondents would be asked four or five of the questions (i.e., only a minority will have incomes exceeding $80,000). The majority would only be asked one, two, or three of the questions in this step-ladder sequence. This step-ladder sequence will lead to far fewer missing income values than a single income question that presents to the respondent essentially the same income categories all at once. It also will yield data that the researchers can combine to form a single income variable with the desired six categories in the original one-question income example.

Step-ladder questions can be used for other constructs that are measured on some form of numerically ordered scale. They are particularly useful when there are many response choices on the response scale and the cognitive burden on many respondents is too great to present all the choices at once. If all choices were presented at once, primacy, recency, and/or other satisficing effects would likely lead to errors in the data.

Paul J. Lavrakas

See also Imputation; Measurement Error; Missing Data; Primacy Effect; Recency Effect; Respondent Burden; Satisficing

STRATA

Strata in stratified sampling are distinct subsets of all entries on a sampling frame. These subsets are strategically defined by one or more stratification variables to improve the statistical quality of findings for the population as a whole or for important population subgroups. Once strata are formed and the sample is allocated among strata, stratified sampling is accomplished by randomly selecting a sample within *each* stratum.

Sampling strata may be formed explicitly from a list of population members or as a part of the selection process in individual stages of a cluster sample. Strata may also be considered "implicit" when the frame is sorted by variables that could otherwise define strata, and sampling is done in each of a set of groups of neighboring entries throughout the frame (e.g., as in systematic sampling). For simplicity, this entry focuses largely on explicit stratification of individual population members, although the same principles of stratum formation apply to other forms of stratified sampling as well.

Stratification Variables

The final set of strata used for sampling is usually some type of cross-classification of the stratification variables. For example, stratification of individuals by two gender categories (i.e., male or female) and four educational attainment categories (i.e., less than a high school diploma, high school diploma, some college, or college degree and beyond) would imply $2 \times 4 = 8$ fully cross-classified strata. If the number of males in the population with some college or a college degree and beyond were considered to be too small, a partially collapsed set of seven strata might be formed from these two stratification variables by using all four female strata and by considering all males with some college or a college degree and beyond to comprise one stratum.

The choice of stratification variables depends on how sample stratification is to be used. When the purpose of stratification is to reduce the variance of the estimate of a characteristic of the population as a whole, it is best for the stratification variables to be statistically correlated with key outcome variables (i.e., member measurements that define the characteristics of interest). Thus, for example, if the outcome variable, annual income, is correlated with educational attainment among population members, then stratification by educational attainment would be a good choice when the population characteristic to be estimated is the average annual income of all population members.

Stratification may also be used to improve the quality of estimates for population subgroups. For instance, when stratification is used to disproportionately sample the subgroup, it is preferable to define strata by whatever characteristics define the subgroup. For instance, educational attainment, as defined earlier, should be a stratification variable when disproportionately sampling persons with less than a high school diploma. Note however, that equally valid though slightly less precise estimates can be obtained for subgroups, even if the categorical variables that would isolate the subgroup are not used to form the strata.

Defining Useful Strata

The ultimate goal in forming strata to improve the quality of estimates for the population as a whole is to define strata that are as internally homogeneous as possible with respect to the survey's key outcome variables. Thus, for example, if the main purpose of a sample of persons is to estimate their average annual income, then one hopes to sample from strata where members of each stratum have relatively similar income levels. This would mean that some strata have low-income earners, others have high-income earners, and the rest separately include those with various intermediate income levels.

In addition to using variables that are highly correlated with key study variables, there are two other considerations in forming strata that will lead to more precise overall population estimates. One is to define "optimum" stratum boundaries when stratification variables are continuous (i.e., numerical, such as age in years or annual income in dollars). The most common approach for doing this is the "$cum\sqrt{f(x)}$ rule," where $f(x)$ is the frequency of the value (x) of a continuous stratification variable among all population members. Optimum boundaries for H strata under this rule are located at points of equal increment in the cumulative of $\sqrt{f(x)}$ (i.e., the $cum\sqrt{f(x)}$) among all values of x.

Another consideration is the number of strata. The general guidance here is that while increasing the number of strata will never diminish the quality of estimates, having too many strata may be impractical or of limited additional benefit. Because a sample size of at least two is needed for unbiased variance estimates, the maximum practical number of strata is the sample size divided by two, as seen in stratified sampling with "paired selection" (i.e., sample size of two in each stratum). Six to 10 well-formed strata are often sufficient to stratify individual population members, while paired selection from equal-sized strata is commonly used to choose clusters in the first stage of multi-stage samples.

William D. Kalsbeek

See also Correlation; Disproportionate Allocation to Strata; Multi-Stage Sample; Stratified Sampling; Subgroup Analysis; Systematic Sampling; Variance Estimation

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: Wiley.
Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Stratified Sampling

Sample selection is said to be stratified if some form of random sampling is separately applied in each of a set of distinct groups formed from all of the entries on the sampling frame from which the sample is to be drawn. By strategically forming these groups, called "strata," stratification becomes a feature of sample designs that can improve the statistical quality of survey estimates. Procedurally, stratum formation and sample allocation to strata are important preliminary steps to sample selection. After the sample has been selected, data are gathered from its members, and analysis accommodating the use of stratification is conducted.

Stratification and sampling clusters are sometimes confused, as both involve groups of one form or another. These two design features are distinguishable by how sampling is applied to the groups. Whereas sampling is done within each of the groups (strata) in stratified samples, only some of the groups (clusters) are randomly selected in cluster samples.

Stratified sampling is statistically beneficial in two ways. First, it may be used to enable the sample to better represent the measurements that define the mean, total, or other population characteristics to be estimated from the sample. A sample also may be stratified to promote adequate sample sizes for analysis of important subgroups of the populations, such as racial/ethnic minority groups.

Forming Strata

Strata are formed by deciding on one or more stratification variables and then defining the actual strata in terms of those variables. Improving the statistical quality of sample estimates through stratification implies that the stratification variables should (a) be statistically correlated with the measurements that define the main population characteristics to be estimated, (b) effectively isolate members of important analysis subgroups, or both. Indeed, the first goal implies that useful strata will be internally homogeneous with respect to the main study measurements, while the second goal suggests that key subgroups should be identifiable by one or more strata. Although the number of formed strata need not be large, it should be large enough to meet the needs of later analysis.

There are two common misconceptions about stratum formation. One is that subgroup analysis will be valid only if the subgroups are defined by those comprising one or more sampling strata. In fact, valid estimates can be produced for subgroups defined by portions of one or more strata, although their precision will be slightly less than if complete strata define the subgroup. Another myth is that incorrect stratum assignment (e.g., due to measurement error in the values of the stratification variables) will invalidate sample estimates. In reality, bias does not result from stratum assignment errors provided random selection is done in each stratum, although the precision that is gained from strata with assignment errors may be compromised by these errors.

Sample Allocation to Strata

Deciding how a stratified sample will be distributed among all strata is called "stratum allocation." The most appropriate allocation approach depends on how stratification is being used. If, for example, the main purpose of stratification is to control sample sizes for important population subgroups, stratum sample sizes should be sufficient to meet precision requirements for subgroup analysis. A special case of this occurs when the subgroups of interest are explicitly defined by individual strata (e.g., stratification by geographic region) and an important part of analysis is to produce comparisons among all subgroup strata. In this instance, equal allocation (i.e., equal sample sizes) among subgroups would be appropriate. Meeting sample size needs for subgroups usually makes the allocation disproportionate, wherein sampling rates differ among strata.

Proportionate allocation is often a prudent choice when the main focus of analysis is characteristics of several subgroups or the population as a whole and where the appropriate allocations for these analyses are discrepant. Proportionate allocation involves applying the same sampling rate to all strata, thus implying that the percent distribution of the selected sample among strata is identical to the corresponding distribution for the population. This allocation is "safe" under these circumstances because, if well-formed strata are used, the precision of resulting estimates from the stratified sample will be at least as good as those from an unstratified sample of the same size.

"Optimum" stratum allocation, in which the most cost-efficient stratum sample sizes are sought, can lead to estimates of overall population characteristics that are statistically superior to those from a proportionately allocated sample. When all stratum unit costs are the same, the stratum sampling rates that yield the most precise sample estimates are directly proportional to the stratum-specific standard deviations of the stratum measurements that define the population characteristic to be estimated. This is called "Neyman allocation." When unit costs vary among strata, the optimum sampling rate for a stratum is also inversely related to the square root of the average cost of adding another sample member to the stratum. A practical limitation of optimum allocation is that its statistical utility is tied to the existence of good measures of stratum standard deviations and unit costs, which may be hard to find.

Sample Selection in Each Stratum

Most stratified samples use the same random selection method (e.g., simple random sampling) in each stratum. Because these selection methods need not be the same in each stratum, stratification offers flexibility that may lead to solutions to certain practical problems facing the survey designer. For instance, using a 5-year-old frame to sample hospitals directly for a national study of patient visits may lead to coverage bias in estimates if a large percentage of new hospitals are left off the list. To solve this problem, a separate stratum of recently opened hospitals might be sampled by first selecting counties and then searching for newly opened hospitals in the selected counties.

Analysis of Stratified Samples

Sample stratification is accommodated in two ways during analysis. First, separate stratum estimates are combined to produce the final estimates. This requires that the analyst be able to identify each respondent by stratum of selection. Second, the approach to stratum allocation is addressed through the use of sample weights that account for any sample disproportionality. These weights may also be calibrated through post-stratification and adjusted to account for other sources of imbalance in sample composition due to nonsampling error.

Failure to properly account for stratification in analysis of stratified samples can seriously invalidate estimated variances of survey estimates and thus compromise analysis findings that rely on them (e.g., confidence intervals and hypothesis tests). More specifically, if variance estimation ignores stratification, the variance estimates will generally be positively biased, thus causing confidence intervals to be too wide and tests of hypothesis to be too conservative.

William D. Kalsbeek

See also Clustering; Confidence Interval; Multi-Stage Sample; Neyman Allocation; Post-Stratification; Random Sampling; Sample Design; Sample Size; Sampling Frame; Simple Random Sample; Strata; Subgroup Analysis; Weighting

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory; Vol. 1. Concepts and discussion*. New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations—Methods and applications*. New York: Wiley.

Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Scheaffer, R. L., Mendenhall, W., & Ott, L. (2006). *Elementary survey sampling* (6th ed.). Belmont, CA: Duxbury.

Straw Polls

Straw polls originated as small, informal public opinion "surveys" that later evolved into large-scale, random sample surveys used primarily to determine the viability of potential political candidates. Nowadays, pollsters offer a litany of straw polls before nearly every national election and several state elections.

The phrase *straw poll* has its origins in the idea of "straws in the wind," which were used to determine which way the wind is blowing. The earliest use of a straw poll was by a newspaper, *The Harrisburg Pennsylvanian*, during the 1824 presidential election. Historically, straw polls were not scientific, and often they were conducted haphazardly. They relied on relatively large and sometimes massive samples to achieve some semblance of accuracy. Early on this included going to public places such as bars, political rallies, train stations and such, where groups of people were gathered and asking them their voting preferences. There was no sampling science at work at all.

However, as straw polling moved into the 20th century, this situation began to change. Yet, even with more than 2 million potential voters surveyed, straw polls were often quite biased, as demonstrated in the often-cited example of the *Literary Digest*'s 1936 declaration that, by a landslide, Alf Landon was going to beat Franklin Delano Roosevelt. (It should be noted that the *Literary Digest* accurately predicted each presidential election from 1920 through 1932.) Recognizing the folly of the *Literary Digest*'s approach, George Gallup, Elmo Roper, and Archibald Crossley each began using a more scientific approach. Specifically, with the help of demographers, they used quota sampling to build their list of poll respondents. Gallup's mistake in 1948 caused the major polling organizations to reconsider that approach. Recognizing the downfalls of a quota sample, subsequent polling adopted much more strict probability methods. These probability methods have been modified to fit specific needs, but they have all held true to one standard: that of random samples.

Straw polls have become commonplace in the news today. They have helped to create the horse race approach to reporting on elections. For example, a Gallup straw poll, conducted December 11–14, 2006, of Republicans and Republican leaders nationwide asked the following: *Next, I'm going to read a list of people who may be running in the Republican primary for president in the next election. After I read all the names, please tell me which of those candidates you would be most likely to support for the Republican nomination for the president in the year 2008, or if you would support someone else?* With an *n* of 425 and a margin of error of ± 6, Gallup found that 28% would support both John McCain and Rudy Giuliani. Another 12% would support Condoleezza Rice, and 8% supported Newt Gingrich.

Although the 2008 Republican nomination process would not formally begin for more than a year after this poll was conducted, the results indicated relatively little support for a number of other potential Republican presidential candidates which might well have caused one or more of them to reconsider actually running. A poll like this also may influence potential funders for particular candidates. It also may help the potential candidates themselves refine their campaign to try to capture support among the population that was not supporting them.

With the availability of the Internet, unscientific straw polls have expanded well beyond the realm of politics. They have been used to vote for favorite contestants on reality television shows, possible actors and actresses for television shows and movies, and even ranking of teams in college sports.

James W. Stoutenborough

See also Gallup, George; Gallup Poll; Horse Race Journalism; Margin of Error (MOE); *n*; Poll; Pollster; Probability Sample; Quota Sampling; Random Sampling; Trial Heat Question

Further Readings

Herbst, S. (1995). *Numbered voices: How opinion polling has shaped American Politics*. Chicago: University of Chicago Press.

Warren, K. F. (2001). *In defense of public opinion polling*. Boulder, CO: Westview Press.

Young, M. L. (Ed.). (1990). *The classics of polling*. Metuchen, NJ: Scarecrow Press.

Subgroup Analysis

Subgroup analysis involves subdividing respondents in a survey into groups on the basis of demographic characteristics (e.g., race, ethnicity, age, education, or gender) or other variables (e.g., party identification, health insurance status, or attitudes toward the death penalty). Analyses of subgroups can be done for a number of reasons. A researcher might analyze differences in variable means or distributions across subgroups to identify disparities or other differences. For example, a researcher studying health care insurance may want to test whether there are differences in the proportion of respondents in different income, education, or race subgroups who are covered by private health care insurance.

Researchers may also want to compare bivariate relationships or multivariate analyses across subgroups to test whether relationships between variables are moderated by subgroup membership. Alternatively, these subgroup comparisons may be conducted to test or establish the generalizability of a relationship or model across subgroups of respondents. For example, a researcher may want to compare the extent to which respondents' pre-election reports of their presidential candidate preferences correspond to their post-election reports of vote choice for respondents who are strong partisans versus those who are weak partisans.

In research involving experimentally manipulated variables, researchers may compare the characteristics of respondents assigned to experimental groups to test whether random assignment has been successful. Researchers also may compare the effect of an experimental manipulation across subgroups to determine if characteristics of respondents moderate the effect of the experimental manipulation or to test or establish the generalizability of the effect across subgroups. For example, survey researchers studying the effects of question order on survey responses could compare the effect of question order for respondents with different levels of education.

For research involving data collected at multiple times (either from the same or different samples of respondents as in longitudinal and panel surveys), subgroup analysis can be used to test whether variable changes over time are the same or different across subgroups. For example, researchers using a panel design to examine changes in math skills in children between the ages of 6 and 10 might compare changes across these ages separately for male and female children.

Subgroup analysis is often used to better understand survey data. However, researchers who intend to use subgroup analysis should keep in mind a number of statistical cautions when using the approach. First, researchers should plan to conduct subgroup analysis in advance (i.e., a priori) rather than deciding to do so after the fact (i.e., post hoc). This helps to address possible concerns with sample size and power. If one or more subgroups are very small, the power to detect effects may be very small, and Type II errors (i.e., concluding there is no difference between groups when there actually is) may be likely. In addition, researchers should be concerned when they are making many subgroup comparisons. Conducting multiple statistical comparisons with the same data increases the chance of Type I error (i.e., concluding there is a difference between groups when the difference is likely due to chance) and researchers conducting subgroup analyses should utilize *familywise error* in estimating the significance of their statistical tests to adjust for this possibility.

Allyson Holbrook

See also Demographic Measure; *n*; Sample Size; Type I Error; Type II Error

Further Readings

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. Hoboken, NJ: Wiley.

Davis, J. A., & Smith, T. W. (1991). *The NORC General Social Survey: A user's guide*. Newbury Park, CA: Sage.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58–88.

Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*. Hoboken, NJ: Wiley.

SUDAAN

SUDAAN is a statistical software package for the analysis of correlated data, including but not limited to the analysis of correlated data encountered in complex sample surveys. Correlation between observations may exist, for example, in multi-stage studies where units such as geographic areas, schools, telephone area codes, or hospitals are sampled at the first stage of selection. A correlation may exist between observations due to the increased similarity of the observations located in the same cluster compared to observations between clusters. During any statistical analysis with survey data, the complex design features of the study, including clustering, unequal weighting, stratification, and selection with or without replacement, should be accounted for during the analysis. Standard software packages that do not account for the complex design features can produce biased results. In fact, in most instances the precision of statistics will likely be underestimated if one were to use a standard software package that does not account for the complex design features of a study. Software packages, such as SUDAAN, that are specifically designed for the analysis of data from survey studies will properly account for many of the complex design features of the study during an analysis.

SUDAAN originated in 1972 at Research Triangle Institute (RTI) and, over the years, has evolved into an internationally recognized, leading software package for analyzing correlated and complex survey data. SUDAAN 9.0, which was launched in August 2004, offers several additions and enhancements to prior releases. RTI statisticians are currently working on the next major release of SUDAAN and plan to have that available in 2008. SUDAAN can be run on any DOS or Windows platform and is also available for the SUN/Solaris and LINUX operating systems (32- and 64-bit operating systems). SUDAAN is available as a stand-alone software package or can be used in SAS-callable format. The SAS-callable format is of particular value to SAS users because it allows SUDAAN to be run directly within any existing SAS job. SUDAAN's syntax is very similar to SAS (e.g., a semicolon must appear at the end of every statement), and SUDAAN has the ability to read data files in various formats such as SAS, SAS XPORT, SPSS, and ASCII.

As noted earlier in this entry, SUDAAN provides estimates that correctly account for complex design features of a survey. These features include the following:

- Unequally weighted or unweighted data
- Stratification
- With- or without-replacement designs
- Multi-stage and cluster designs
- Repeated measures
- General cluster-correlation (e.g., correlation due to multiple measures taken from patients)
- Multiply imputed analysis variables

Currently, SUDAAN is a single program that offers nine analytic procedures. SUDAAN is one of the few software packages that offers three of the most commonly used robust variance estimation methods, including Taylor series linearization (generalized estimating equations for regression models), jackknife (with or without user-specified replicate weights), and balanced repeated replication.

All three variance estimation methods are available in all of SUDAAN's procedures. The nine procedures consist of three descriptive procedures, four regression procedures, and two survival procedures.

Michael B. Witt

See also Balanced Repeated Replication (BRR); Clustering; Complex Sample Surveys; Jackknife Variance Estimation; Multi-Stage Sample; Sampling Without Replacement; SAS; Taylor Series Linearization

Further Readings

SUDAAN: http://www.rti.org/sudaan

Suffix Banks

Telephone numbers in the United States, Canada, and the Caribbean consist of 10 digits divided into three components: The first three digits are the area code; the next three digits are the prefix or exchange; and the final four digits are the suffix, or local number. For each area code-prefix combination, the 10,000 possible numbers for a suffix can be subdivided into banks or blocks of consecutive numbers: 1000-banks (Nnnn), 100-banks (NNnn), or 10-banks (NNNn).

As the sampling method of random-digit dialing (RDD) evolved, the underlying assumption was that residential telephone numbers tended to cluster. Based on this assumption, one of the earliest methodologies consisted of simply adding one to a directory-listed number or randomizing the last digit of a directory-listed number. Both methodologies proved to introduce bias. At the other end of the spectrum, randomizing all four digits of suffixes in residential exchanges eliminated bias but proved to be too inefficient and expensive to field.

Since the early mechanical switches came in banks of 1,000 numbers, telephone companies avoided purchasing unnecessary equipment by assigning numbers in suffix blocks of 1,000 numbers (or "1000-banks"). Later research showed that this clustering of numbers extended to 100-banks as well, primarily to accommodate businesses.

Two classes of RDD methodologies were developed to take advantage of this clustering effect: two-stage designs (such as Mitofsky-Waksberg) and single-stage, list-assigned designs that use a database of listed numbers to qualify banks (1000-banks or 100-banks) for inclusion in the telephone frame, based on the presence of directory-listed numbers. Although not all countries have fixed-length telephone numbers, these designs have been successfully adapted for use around the world.

Linda Piekarski

See also Cell Phone Sampling; Mitofsky-Waksberg Sampling; Prefix; Random-Digit Dialing (RDD)

Further Readings

Lepkowski, J. M. (1988). Telephone sampling methods in the United States. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 73–98). New York: Wiley.

Piekarski, L. (1996). *A brief history of telephone sampling*. Fairfield, CT: Survey Sampling, Inc.

Tucker, C., Lepkowski, J. M., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly*, *66*, 321–338.

SUGing

SUGing, or Selling Under the Guise of research (also known as "sugging"), is the act of telemarketing or pre-qualifying customers while pretending to conduct a legitimate survey. Because it is, by definition, a practice that seeks to deceive the respondent, it is an unethical business and research practice. In addition, there is essentially no interest in the "data" being gathered other than to help make the sales pitch more effective.

It also has the side effect of increasing the victim's suspicion of subsequent legitimate survey contacts, thereby reducing the population of individuals readily willing to answer survey questions. In short, SUGing not only represents illegitimate and thus unethical research in itself, but it also has a long-term negative impact on legitimate research projects. SUGing is proscribed by numerous public opinion, market research, and direct marketing trade associations.

In SUGing, the respondent's answers to the purported survey questions serve as a means of setting up a sales pitch. This is in contrast with legitimate survey research, in which obtaining the respondent's answers is the desired goal. Questions used in SUGing, therefore, are often superficial or biased to create a favorable disposition toward the sales solicitation. In some cases, however, SUGing solicitations can be extremely detailed in order to develop a detailed profile of the respondent to facilitate a custom tailored sales pitch that arrives later (e.g., via mail), which itself may represent another violation of ethical obligations to protect respondent confidentiality. Regardless of the nature of the questions, a key feature of a SUGing solicitation is its deceptiveness. There is no indication of the true nature of the interview until the sales pitch is given.

SUGing is illegal in several jurisdictions, including the United States, Canada, and the European Union. As a complicating factor, though, there is some international disagreement about the exact meaning of *SUGing*. Although public opinion and market research organizations in the United States, the United Kingdom, and continental Europe all define *SUGing* as selling under the guise of research, in Canada the *S* also stands for "soliciting" (giving the term a similar meaning to *FRUGing*, which is fund-raising under the guise of research), and deceptive survey-like sales pitches are known as *MUGing* (marketing under the guise of research).

One area where legitimate survey research may seem like SUGing is the practice of "sales waves." A sales wave attempts to determine the sales potential of a new product by offering to sell the new product to a respondent immediately at, or soon after, the conclusion of a legitimate market research interview about the product. Although the intent of a sales wave is to understand product demand and not to sell the product, the structure of a sales wave is similar enough to SUGing as to potentially create misunderstanding and resentment among many respondents.

The extent of SUGing is difficult to estimate, although empirical research on nonresponse in North America has indicated that up to 50% of individuals have received a SUGing call or mailing.

Geoffrey R. Urland and Kevin B. Raines

See also Deception; Ethical Principles; FRUGing; Nonresponse; Pseudo-Polls; Survey Ethics; Telemarketing

Further Readings

Kimmel, A. J. (2007). *Ethical issues in behavioral research: Basic and applied perspectives*. Malden, MA: Blackwell.

Market Research Association. (2007). *Research resources: For the consumer—Research abuse (sugging and frugging)*. Retrieved February 15, 2007, from http://www.mra-net.org/resources/abuse.cfm

Superpopulation

When data for a variable are gathered from a finite population and that variable is regarded to be a random variable, then the finite population is referred to as being "a realization from a superpopulation." A superpopulation is the infinite population that elementary statistical textbooks often describe as part of the enumeration of a finite population. It is because sampling theory is based on making inference for a well-defined finite population that the concept of superpopulation is needed to differentiate between a finite population and an infinite superpopulation.

This distinction is important for two reasons: (1) Sampling theory estimation and inference can be based entirely on a finite population (in the absence of nonsampling errors), with no recourse to a superpopulation; and (2) even when a superpopulation is of primary interest (such as parameter estimation of the superpopulation model), the finite population may have been sampled in a way that distorts the original distribution of the finite population.

The superpopulation and the finite population concepts are compatible if one views the finite population labels (which are needed to allow specific units to be sampled) to be part of the superpopulation model. Doing so, the final sample can be thought of as the result of a two-step process. First, the finite population is selected from a superpopulation according to the superpopulation model. Then, after each unit is identified with a label and related information, a sample design is formed and the final sample of units is selected. The measured characteristics are known only for the final sample. However, other information, such as the information used as part of the sample design, will be known for the entire finite population.

The superpopulation approach allows the use of additional model assumptions by specifying either a frequency distribution for the finite population characteristics or by specifying a prior distribution directly for them. Including this extra information as part of the inference often increases precision. A potential danger is that inference may be either biased, due to model misspecification, or inappropriate if prior information used is not shared by others.

A different but related concept is that of a nested sequence of populations that increases to an arbitrary large total. This has been used to demonstrate asymptotic properties of finite population estimates.

Donald J. Malec

See also Finite Population; Inference

Further Readings

Cassel, C.-M., Särndal, C.-E., & Wretman, J, H. (1977). *Foundations of inference in survey sampling*. New York: Wiley.

Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, *19*(4), 535–545.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys*. New York: Wiley.

Supervisor

Survey research supervisors are tasked with two main responsibilities: managerial or administrative duties and research-specific duties. It is the supervisors who are responsible for ensuring projects are proceeding on schedule and are within budget, the research process is carried out with due regard for ethical obligations, and all research goals are satisfied. To do this, supervisors must have a firm grasp on the operations of their research firm, knowledge of the entire research process, and an understanding of business policies and practices. A comprehensive knowledge of the science of survey research is also very helpful to supervisors. In addition, direct experience with various survey tasks is an inherent necessity for survey research supervisors.

Administrative Duties

Research operation supervisors typically oversee a staff of department or area supervisors, administrative personnel, interviewers, and other research assistants. Department supervisors are often responsible for specific tasks, such as data collection, computer-assisted data collection programming, or data coding and editing. These supervisors are responsible for the day-to-day supervisory activities.

Proposals

Research supervisors are often tasked with the responsibility of helping to submit proposals. This can become an arduous activity with little promise of success. Typically for large contracts, multiple proposals for different projects must be submitted before a project is approved. It is beneficial for the supervisor to have a full-time staff member assisting with proposal submissions.

Budgetary

Budget is a primary concern for all organizations. Supervisors must have a clear picture of all operational costs affecting their department, including facility costs, payroll, and supply expenditures. Research supervisors may be responsible for estimating project costs and then ensuring actual costs fall within those projections.

Hiring

Supervisors are engaged in various aspects of the hiring process. This may include advertisement, application review, interviews with candidates, and often acting as the liaison between human resources and their data collection staff.

Staff Performance Evaluation

Performance evaluations are critical to the growth of an organization. Evaluations outline goals for individuals which, if carried out, can improve the organization's performance. It is imperative supervisors have in place an equitable and regular evaluation system.

Research Duties

Project Development

Research supervisors are often tasked with developing the operational features of projects from the ground up. They may work in cooperation with their client or with topical experts, as necessary, to develop the operationalization of the research design, questionnaire, and data processing stages. Of course, this varies greatly depending on the specific research firm or project. At times, research supervisors may be fulfilling only specific research requests such as data collection. This is often the case when research firms conduct data collection for government agencies.

Survey Design

Clear research objectives need to be set in advance of creating any research design. The target population and points of interest must be outlined. Question design is both a science and an art form and should be handled by the most experienced and knowledgeable people. Objectivity should be the goal of every question. This can be difficult when researching heated or political topics. Supervisors have the responsibility to make certain all these aspects of the survey design have been addressed and resolved before a project can begin data collection.

Sampling and Sample Size

Supervisors are responsible for the successful processing of the sample. This includes issuing new cases to interviewers on a timely basis, overseeing the system that sets callbacks and recontacts, sending out follow-up mailings; and making certain that the final sample size of completed interviews is accomplished within the needed time frame and, ideally, on budget. If problems arise during the field period, especially ones that were unanticipated by the researchers, it is the responsibility of supervisors to bring these problems immediately to the attention of their superiors and seek their immediate resolution. They must ensure that the solution is implemented properly, and they must monitor the effectiveness of the solution and provide continuous feedback to their superiors.

Timelines

Supervisors must ensure a realistic schedule is set at the beginning of the project and that this schedule is followed. Regular meetings with department supervisors provide up-to-date information on a project's progress. The research process seems to always have its share of troubles which inherently affects the calendar. When available, a buffer window should be built into the schedule to protect against unforeseen difficulties, and it is the supervisors' responsibility to see that this happens.

Interviewer Training and Monitoring

Developing, implementing, and maintaining a comprehensive interviewer training and monitoring system is a central task of research supervisors. Training and monitoring systems build flexibility into an organization in case unexpected absences or needs arise. Often it is the task of department supervisors to develop training materials for their respective research areas.

Jeff Toor

See also Interviewer Monitoring; Interviewer Training; Mode of Data Collection; Questionnaire Design; Research Design; Research Management; Sample Design; Sample Size; Supervisor-to-Interviewer Ratio; Survey Costs; Survey Ethics

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Rea, L., & Parker, L. (2005). *Designing and conducting survey research: A comprehensive guide*. San Francisco: Jossey-Bass.

Stouthamer-Loeber, M., & Bok van Kammen, W. (1995). *Data collection and management: A practical guide*. Thousand Oaks, CA: Sage.

SUPERVISOR-TO-INTERVIEWER RATIO

This entry addresses the issue of how the average number of interviewers that are assigned to a survey supervisor affects data quality. There are many factors that affect survey interviewer performance and success. Individual skill set, training, experience, and questionnaire design all influence the ability of interviewers to gain cooperation and gather accurate data. The one ongoing factor that has a significant impact on data quality in surveys that are interviewer-administered is the interaction of the supervisor with the telephone interviewer. Sometimes called a monitor, coach, team lead, or project/section/floor supervisor, this staff position gives management the greatest leverage to influence the human aspect of such survey projects.

The supervisors of survey interviewers may fulfill many roles in a survey project, including any or all of the following:

- Monitor the quality of the data being collected.
- Motivate the interviewers to handle a very difficult job.
- Provide insight into how a study is going and what the problem areas are.
- Provide training to improve the interviewing skills of the interviewers.
- Supply input for the interviewer performance evaluation.

It is the supervisor who is in daily contact with the interviewers and who knows each interviewer's strengths and skills to be developed, maintained, or enhanced. The supervisor answers questions, guides interviewers through new or difficult processes, and sets the mood, pace, and outlook of the whole team. A supervisor who is positive, uplifting, professional, and supportive will increase retention of interviewers, improve data quality, and lower costs of the project. There have been many human resource studies that show employees are most influenced by their immediate supervisor to be happy in their jobs or to leave a job. To ensure the best results on survey projects, research management must not only select the right people to be motivating, inspiring supervisors but also determine the right ratio of interviewer to supervisor.

The right ratio for a project is the one that provides enough supervisory coverage to meet the above goals: retain interviewers, ensure good quality, and provide cost efficiencies. A ratio that is too high (e.g., > 1:20) for a project's needs will lead to too little interaction with interviewers, lower-quality data, and increased turnover. A ratio that is too low (e.g., 1:5 or less) for a project's needs will increase costs and may lower supervisory motivation due to boredom. To meet the challenge of determining the correct ratio for a project, research management can use a checklist of criteria questions.

Some suggested questions to help develop the right ratio include the following:

1. Is the study a new type of survey in its content, group to be sampled, procedures, or client/industry? If the interviewers are not experienced with the type of study being done, the supervisor-to-interviewer ratio should be lower than the normal benchmark (e.g., < 1:10).

2. Is the study an ongoing survey that will change very little, or will the questions, procedure, shifts, and sample change frequently? An ongoing study with few changes will need a lower ratio of supervisor to interviewer, but a survey with constantly changing requirements needs proportionally more supervision.

3. Does the study have very stringent or complex requirements, such as complicated sample quotas, sensitive or difficult selection criteria, or comprehensive validation? If there are other measurements besides real-time monitoring of interviewers' work and answering questions that must be tracked hourly, daily, or frequently within a shift by the supervisors, such as in the case of centralized telephone surveys, then a lower ratio is demanded.

4. Is the project a business-to-business study, or a social/political study that requires respondents who are professionals or executives, like doctors, lawyers, chief executive officers, and so forth? Interviewers will have to possess more professional skills on these projects and may need more guidance from supervisors on getting through gatekeepers and other obstacles. Surveys conducted with professionals or highly

affluent households or households with high-security access challenges often require a lower ratio than general consumer surveys.

5. Does an ongoing project typically have a high percentage of new or inexperienced interviewers? If a survey organization experiences a steady, high level of turnover and an influx of raw interviewers, the supervisor-to-interviewer ratio will need to be lower to support the development of this staff.
6. Is the study or project one that is audited by an independent group, or is there a third party that has oversight on the survey procedures and requirements, such as a government agency or industry watchdog committee? The project may need proportionally more supervisors to ensure compliance with the audit or oversight.
7. Does the client funding the project represent a major portion of the survey organization's business or is otherwise a high-profile client with heavy impact on the organization's success? If so, then management may allocate more supervision to make sure all aspects of the work are going well and according to client expectations.
8. What is the budget for the project? How was the bid costed out when the work was contracted? How much supervisory costs were built in? The manager may be restricted in what he or she can work with on the ratio by the cost expectations already set up by the company or the client.

Once all of these questions have been answered, the project management can derive the supervisory ratio best suited for that project. There has to be a beginning point, or benchmark, that the manager uses to develop the specific ratio. Although there is no one benchmark used by all survey research data collection centers, a common range of ratios for supervisor to interviewer in many commercial organizations that do interviewing is in the range of 1:15 to 1:20 of supervisor per interviewers. Starting with that range, the manager can work up or down in setting the ratio against the sample questions listed previously. In contrast, since the 1980s, Paul J. Lavrakas has advised that for high-quality supervision, this ratio for supervising telephone interviews should be more in the 1:8 to 1:10 range.

As an example, imagine an ongoing general consumer survey that changes only every 6 months and uses 60% or more experienced interviewers. This survey will probably be fine with a ratio of one supervisor to 15 interviewers (or more). In contrast, a new study on a complex technical issue, where all the interviewers are new and the respondent pool is sensitive, might require a ratio of one supervisor to 8 interviewers in the very beginning or if the study has short duration. Or, for example, an ongoing study that adds or changes questions every 3 or 4 days, that has complex sample quotas, and that has 40% experienced interviewer levels may well require a ratio of one supervisor to 10 to 12 interviewers.

The data collection manager can experiment with the ratio at the beginning of the project, observing the interaction of supervisors with interviewers, checking the data quality, and watching for problems and issues at first, to decide if the ratio is too high, too low, or just right. It is better to start out with a lower ratio (thus a more conservative, but also more expensive, ratio) if the manager has no experience with a particular set of survey conditions, and then raise the ratio (i.e., more interviewers per supervisor) if appropriate as the project progresses.

If still unsure after using the previously listed criteria questions, a research manager can consult other managers in the organization, seek out peers in the industry through association contact, or look for information in research industry association publications, guidelines, or training material. There are also courses on project management and field director management that a research manager can participate in for better understanding.

In the end, the research manager should rely on his or her own experiences, instincts about the project, and "eyeball" observations of the daily activity when the study first starts. Valuable information also can be obtained by asking both the supervisors and interviewers on the study for their thoughts on how the supervision is working. Every study will have its own dynamics as far as the supervisor-to-interviewer ratio is concerned; the wise manager will learn from past experience and from peer input what works best on various kinds of studies.

Kathy Pilhuj

See also Interviewer; Interviewer Monitoring; Supervisor; Validation

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

SURVEY

Survey is a ubiquitous term that is used and understood differently depending on the context. Its definition is further complicated because it is used interchangeably as a synonym for other topics and activities listed under the broad classification of survey research or survey methods. There are multiple uses of the term *survey* that are relevant to particular linguistic applications. Survey is used as a noun when it refers to a document (e.g., fill out this survey) or a process (e.g., to conduct a survey); as an adjective (e.g., to use survey methods); or as a verb (e.g., to survey a group of people). *Survey* is used interchangeably with and strongly associated with the terms *poll* and *public opinion polling*, and survey methods are used to conduct a census—an enumeration of the total population. In addition to being used to identify a type of research tool, survey research is a subject that can be studied in an educational course or workshop, and it is an academic discipline for undergraduate and graduate degrees. This entry focuses on the basic definition from which these other uses of the term *survey* have evolved by outlining the essential elements of this term as it relates to the scientific study of people.

A survey is a research method used by social scientists (e.g., economists, political scientists, psychologists, and sociologists) to empirically and scientifically study and provide information about people and social phenomena. A survey is scientific because there is an established process that can be followed, documented, and replicated. This process is rigorous and systematic. The typical steps in the survey process are (a) problem formation, (b) hypothesis development, (c) research design, (d) sample design and selection, (e) questionnaire development, (f) data collection, (g) data analysis, (h) reporting and dissemination, and (i) application of information. Underscoring the complexity of a survey, each of these steps in the process also has a set of essential and accepted practices that is followed, documented, and replicated, and specific professional training is required to learn these practices. The documentation that accompanies a survey provides the information necessary to evaluate the survey results and to expand the understanding and analysis of the information provided from the survey. Sampling error and nonsampling error are two dimensions of survey error. Most scientific surveys could more accurately be called "sample surveys" because probability theory is used to scientifically select subgroups of the population to study, and there is a body of knowledge on acceptable statistical procedures for sampling and for calculating sampling error. Quantifying the nonsampling error associated with other steps in the survey process (e.g., question wording, interviewer effects, and item and unit nonresponse) is more challenging. Total survey error incorporates the error possible in any of the steps in the survey process. When the results of a survey are reported and, particularly, when people use survey results to make decisions, it is very important to review the documentation that describes how the survey was conducted so the quality of the data can be assessed.

A survey can be used to find out the opinions, attitudes, and behaviors of persons who are contacted to participate in the survey and to obtain other factual information about members of this population. (Surveys also can be conducted to study animals other than humans, such as buffalo in a game preserve; crops in a field; or inanimate objects, such as books in a library collection.) The information from individuals is then aggregated to provide a statistical profile of the survey population. Surveys are conducted by many types of organizations and researchers. Federal, state, and local government surveys are conducted to obtain information to guide public policy decisions, and some surveys are legislatively mandated to evaluate social programs. A well-known government survey is the U.S. Census, which attempts to collect information about every person in the United States. More typically, the government conducts sample surveys to learn about and monitor topics such as employment trends (Bureau of Labor Statistics) and health issues (Centers for Disease Control and Prevention and the National Center for Health Statistics). Surveys are used by academic researchers to test hypotheses such as those related to social behaviors (e.g., marriage and families, alcohol and drug consumption, preparation for retirement) and to conduct social experiments (e.g., cost-effectiveness of different interventions to prevent obesity). Corporations use surveys to make decisions about the products they invest in and bring into the marketplace and to determine customer satisfaction with these products after they have been purchased. Familiar to many households is the Nielsen TV Ratings survey, which monitors the public's use of television. The Gallup Poll and the Pew Research Center, as well as electronic and print news organizations (e.g., *New York Times*/CBS News poll; ABC News/*Washington Post* poll) use surveys to provide timely profiles and to track public opinion about current issues. A prominent use of surveys is the pre-election polls that inform candidates and the public about important campaign issues.

Janice Ballou

See also Complex Sample Surveys; Poll; Research Design; Response Rates; Standard Definitions; Total Survey Error (TSE)

Further Readings

American Statistical Association. (n.d.). *What is a survey?* Retrieved April 19, 2008, from http://www.whatisasurvey.info

Converse, J. M. (1987). *Survey research in the United States: Roots and emergence 1890–1960*. Berkeley: University of California Press.

Turner, C. F., & Martin, E. (Eds.). (1984). *Survey subjective phenomena* (Vol. 1). New York: Russell Sage Foundation.

Survey Costs

Survey costing is a complex process that balances a survey organization's financial objectives against the expenses associated with achieving or maintaining the scientific standards that govern validity and reliability, or quality, of the final product. Achieving optimum balance between budgetary and scientific goals requires that researchers first understand how survey operational components are related to costs and how changing each influences both data quality and budgetary outcomes.

It is important to separate survey costs (direct and indirect, variable and fixed) from survey price, as price includes costs plus profit, for those organizations that are for-profit. Profit does not necessarily have any relationship to cost other than being added to it to create total price. As a result, profit and price are not discussed in this entry.

Managerial Accounting

Managerial accounting defines the principles used to develop costs. It differs from financial accounting in that it is intended to provide insight for better organizational decision making. As such, managerial accounting has no standards, and its formulas and data are often unique to an organization. To learn more about managerial accounting, see the Institute of Management Accountants' Web site (www.imanet.org).

Much of what goes into managerial accounting is built on an organization's operations and is often proprietary and confidential. However, there are some basic concepts that are universal to costing in all survey operations. These concepts, and the costing principles derived from them, provide the foundation for any organization to understand and build its own managerial accounting system for survey costing and budget management. Understanding them helps researchers understand how their budget and costs are related to survey quality.

Managerial accounting uses six basic concepts to define costs:

1. Cost object: The thing managers want to know how to cost
2. Cost driver: Any factor that affects costs
3. Variable cost: A cost that changes in proportion to a *cost driver*
4. Fixed cost: A cost that does not change in total
5. Direct cost: Those costs traced to a cost object. These are usually, but not always, variable costs.
6. Indirect cost: Those costs allocated to a cost object. These are often, but not always, fixed costs.

What constitutes a fixed or variable cost changes by survey mode. What constitutes a direct or indirect cost can also change by mode of data collection, but usually less often. What is important to remember is that the determination is made by the cost drivers and object(s).

Some cost drivers are generally constant across all survey modes. General cost drivers that cross survey modes include the following:

- Number of completed interviews (*n* size)
- Topic
- Survey length
- Time in field
- Data manipulation
- Validation requirements
- Net effective incidence
- Desired response rate

- Analytical labor
- Travel to client

For example, if the cost object is a computer-assisted telephone interviewing (CATI) survey, cost drivers might include the following list of items:

- Length and complexity of CATI programming
- Number of open-ended questions and associated coding
- Desired data outputs and associated reporting

Once the cost drivers have been identified for a particular cost object, a list of direct and indirect costs associated with those cost drivers can be constructed. These cost categories will also usually be constant across organizations, although the actual costs associated with each can be highly variable. Using the CATI example, an organization might see the following direct and indirect costs:

Direct Costs:

- Interviewer labor
- Telephone long distance
- Production floor management
- CATI programming
- Cost of sample
- Data manipulation
- Manual reports

Indirect Costs:

- Telephone line charges
- Management
- Rent
- Insurance
- Technical infrastructure

For many organizations some costs tend to remain fixed regardless of survey size. These could include upper management, rent, insurance, technical infrastructure, manual reports, and data manipulation.

A closer look at the CATI example will demonstrate that the majority of direct and indirect costs for this mode are variable costs, meaning that their amount goes up or down based on the size of the survey sample. This is due primarily to the fluctuations in interviewer labor required. Interviewer and floor management labor, telephone long distance, and power are just a few of the variable costs tied to labor. Other variable costs, including cost of sample and CATI programming, are tied to survey length and sample size.

In contrast, consider an Internet survey. In addition to the general cost drivers outlined earlier, cost drivers for an Internet survey might include:

- Concurrency/bandwidth
- Data manipulation
- Manual reports
- Length and complexity of programming

It is important to remember that in some cases, whether a cost is variable or fixed depends on the organization's managerial accounting approach. Similarly, whether a cost is considered direct or indirect is a function of the accounting approach.

Rate Calculations

Because of the dominance of variable costs in survey data collection modes which rely on interviewing labor (in-person and telephone), formulas to assist in calculating production accuracy have been developed to predict costs for these modes. This is true particularly of survey modes that are applied to difficult-to-reach populations and those with large sample sizes. In these cases, variable direct costs can fluctuate tremendously with the slightest change in specification.

There are three fundamental formulas used to predict cost for surveys that include interviewing labor. These are (1) net effective incidence rate, (2) respondent cooperation rate, and (3) production rate.

These three rates often have different names, depending on the country or even which professional association's standards are being employed within an organization. It is not the intent here to suggest that any one association's or organization's labels are preferred.

Fortunately, the research industry in the United States has reached a point where the mathematical formulas for calculating net effective incidence and respondent cooperation rate are generally agreed upon. Production rates are derived using incidence and respondent cooperation calculations. They are managerial accounting equations and, as such, are proprietary and confidential. However, a universal standard for CATI surveys is offered here as an example and as a baseline for any organization or researcher building an accurate production rate.

Incidence is another way of saying frequency of a desired characteristic occurring in some group of people. It is sometimes expressed as eligibility. There are as many names for incidence as there are types of research. The term *incidence* is business language used primarily in commercial research, though the concept and mathematics involved are universal to all forms of survey research.

What is important to remember from a costing perspective is that there may be a difference between incidence of a desired characteristic within the universe versus among the selected population from which the sample frame is drawn. If the sample frame is poorly drawn, there may yet be another difference between the selected population and the sample frame.

Net effective incidence is one industry term used to mean the percentage of respondents known to be eligible for the survey in the sample frame, meaning that they have been qualified through contact. *Net effective incidence* is the term used by companies whose business language is influenced primarily by the standards of Marketing Research Association (MRA) and the Council of American Survey Research Organizations (CASRO), among others. Eligibility calculations are the equivalent for those organizations whose business language is driven by the American Association for Public Opinion Research (AAPOR) formulas.

Net effective incidence is a critical variable in the production rate equation. It is therefore important to accurately estimate what the net effective incidence rate will be when estimating costs. The calculation for actual net effective incidence is effectively a function of multiplying the percentage of the sample frame eligible for each desired criterion by the percentages for other desired criteria. This can be expressed as

$$Q_1 * Q_2 * Q_3 * Q_4 = \textit{Net Effective Incidence}.$$

It is important to note that for costing purposes net effective incidence is based on actual eligibility rates of screened respondents only. Initial estimates must be based on the estimated or projected eligibility rate of screened respondents. This difference makes knowing in advance the eligibility differences of the desired universe, population, and sample frame design used important to survey costing.

Net effective incidence also can be used to make estimates of screenings that will be required to achieve minimum sample sizes by multiplying incidence by minimum sample size desired. This is sometimes referred to as *net incidence*. It is a useful calculation to survey costing in predicting sample and labor requirements, for example. It is important to note, however, that achieved cooperation rate (discussed later in this entry) also influences how much sample and labor are required. The formula for the net incidence calculation, using a prediction of cooperation rate, would be

$$(Q_1 * Q_2 * Q_3 * Q_4) * \textit{Cooperation Rate} = \textit{Net Incidence}.$$

For those organizations that have response rates as a cost driver, total projected interviewing labor effort to meet a response rate goal can also be derived from this calculation by building on net incidence.

Cooperation rate is a term used widely throughout the research industry and has various definitions based upon which association standards are being used. The use of different definitions will yield different rates from the same formula, so it is important to note which definition is being applied in any production formula.

The MRA defines cooperation rate as the percentage of all qualified respondents who agree to complete an interview. The AAPOR has four definitions, all derived from a definition that the rate is the proportion of all cases interviewed of all eligible units ever contacted. CASRO, the Advertising Research Foundation, and the CMOR (Council on Marketing and Opinion Research) all use the AAPOR definition. The definition used here is that of the proportion of all cases interviewed of all eligible units ever contacted, or the AAPOR COOP1 formula expressed as

$$COOP1 = \frac{\textit{Interviews}}{(\textit{Interviews} + \textit{Partial Interviews} + \textit{Refusals} + \textit{Other})}.$$

Where live interviewing is involved, cooperation rates are influenced by the skill of the interviewing staff, the topic of the survey, the length of the study, and the population being interviewed. As a result, the rates themselves are unique to the operation conducting the work and often proprietary. Cost estimates must be built using cooperation rate data from past performance inside the same operational unit on previous, similar surveys with similar populations. Once a survey is under way, more accurate projections can be made from the survey's own performance for cost control purposes.

Benchmarks are available for some types of general population research. CMOR publishes average industry cooperation rates annually for CATI research.

With cooperation rate and net effective incidence in hand, a researcher or organization can accurately estimate production rate: the holy grail of labor-intensive, variable-cost survey modes. The values for variables in production rate formulas are unique to individual production operations. They are based on management effectiveness, labor quality, operational standards, quality control standards, and even weather. They are also closely guarded to maintain competitiveness.

However, the variables themselves in the production rate formula are universal. A CATI production rate formula is given here as an example.

$$\frac{(60\ minutes - B)}{L + S((1/I) - 1) + W/(C * I) + X}$$

B = break, brief, coach time in minutes (any idle time)

L = length of one survey in minutes (including screener)

S = average time to screen per survey in minutes

W = wait time between contacts in minutes

C = cooperation rate as %

I = net effective incidence as %

X = wrap time per interview in minutes

The nature of the variable labels makes clear the degree to which operational and management factors influence survey costing as much as survey design in projects using live interviewers. Break and coaching time are functions of operations management. Wait time is driven by technology. Survey length is a function of design.

Production rates are designed to derive a number of completed surveys per hour (or in some cases, a number of hours per completed survey). Having a solid estimate or mid-term projection of this figure allows for accurate cost estimating for variable data collection costs. Even in cases where live interviewing labor is not required, one or more rates are often an appropriate method for estimating some variable, direct costs (e.g., cooperation rates for self-administered surveys via mail or Internet to project mail-out rate to receive sufficient returns on schedule).

There are live interviewing environments where these formulas need be modified or even eliminated. Some countries of the world pay CATI or face-to-face interviews by the piece rather than per hour. In these cases, net effective incidence and cooperation rate may become less important cost drivers, affecting direct variable cost of management labor and sample acquisition, for example, but not interviewing labor. The use of these formulas, as with all other costing calculations, is dependent on the cost drivers involved.

Estimating and Controlling Costs

Derived production rates per hour allow organizations and researchers to project estimated or actual variable, direct costs for surveys. To turn those hours into a budget requires the application of costs per hour. How those costs per hour are derived is as variable as there are managerial accounting systems.

Fixed and indirect costs are added as line items to variable and direct costs to build an entire budget. In others they are allocated by the hour. In some cases, these costs are estimated at their actual usage by a specific project. In some cases, they are allocated proportionally to entire labor hours. In yet other cases, they are derived from estimated company expenses which are included in a markup ratio to project expense.

Researchers need to understand how their organizations' accounting rules affect the way their costing is designed in order to control and modify their budgets most effectively. As a study progresses, estimated rates and the budget built on them can be replaced with actual rates to project final costs. When a budget shortfall is projected, researchers can then dig into all the cost drivers that go into the costing equation to find ways to realign costs while balancing quality considerations.

Some common approaches to budget realignment include the following:

- Increase field time, which can boost cooperation rates and boost production.
- Decrease total sample size, screening criteria, or response rate goals to reduce labor and sample required.
- Change the proportionality of samples that are not reflective of the population's natural characteristics, thereby increasing net effective incidence.
- Reduce production quality control standards, thereby reducing floor management labor costs.
- Move work to an operation with less expensive costs.
- Incentivize labor to produce at a more rapid rate.
- Renegotiate price with operations supplier.

Net effective incidence deserves special note here. The relationship between incidence and production rate is not linear, but rather curved. Low incidences have much lower rate of completion per hour than high incidences. As a result, the most effective way to increase production rates in low incidence studies is to eliminate screener questions.

Of course, eliminating screeners, as well as most of the other options listed here has direct and significant impacts on survey validity and reliability. Researchers must use their best judgment and understanding of how costs are derived and controlled to design and manage cost-effective yet scientifically useful research.

Karl G. Feld

See also American Association for Public Opinion Research (AAPOR); Coding; Computer-Assisted Telephone Interviewing (CATI); Cooperation Rate; Council for Marketing and Opinion Research (CMOR); Council of American Survey Research Organizations (CASRO); Eligibility; Interviewer Productivity; Mode of Data Collection; Quality Control; Research Management; Total Design Method (TDM)

Further Readings

American Association for Public Opinion Research: http://www.aapor.org

American Association for Public Opinion Research. (2004). *Standard definitions: Final dispositions of case codes and outcomes rates for surveys*. Lenexa, KS: Author.

Council of American Survey Research Organizations: http://www.casro.org

Council of American Survey Research Organizations/ Marketing Research Association (1995). *Incidence guidelines*. Port Jefferson, NY, & Rocky Hull, CT: MRA-CASRO.

Dawson, K. (2004). *The call center handbook*. San Francisco: CMP Books.

Feld, K., & Haynes, D. (2004, October). *Costing quantitative live interview and Web data collection*. Short course taught for the Marketing Research Association, Richmond, VA.

Institute of Management Accountants: http://www.imanet.org

Marketing Research Association: http://www.mra-net.org

SURVEY ETHICS

Survey ethics encompasses a set of ethical procedures that are intended to guide all survey researchers. These procedures are essential to the research process so that explicit care is taken that (a) no harm is done to any survey respondent, and (b) no survey respondent is unduly pressured or made to feel obligated to participate in a survey. This entry discusses informed consent, the rights of respondents, and the social responsibility of researchers.

Informed Consent

The ethics of survey research and its importance often are not covered thoroughly in research methods textbooks and courses. However, the acquisition of knowledge through survey research requires public trust and, therefore, researchers must adhere to ethical practices and principles involving human subjects. Most governmental bodies throughout the industrialized world have established ethical guidelines for conducting survey research. Research proposals often are subject to regulatory review by one or more ethics boards (e.g., an institutional review board) that have been established to ensure that the ethical guidelines set forth by the governing body will be followed. In addition to regulatory bodies overseeing research activities involving human subjects, most professional organizations and associations also have established guidelines and standards of conduct for conducting research that are expected to be maintained by organizational members. The primary tenet among these governing bodies and organizations, as it relates to carrying out research on human subjects in an ethical manner, is that the researcher be cognizant of research participants' rights and minimize the possibility of risk (i.e., avoid exposing research participants to the possibility of physical or psychological harm, discomfort, or danger).

To that end, central to all research ethics policy is that research participants must give their informed consent voluntarily. The purpose of informed consent is to reasonably ensure that survey respondents understand the nature and the purpose of the survey, what is expected of them if they participate, the expected length of time necessary for them to complete the survey (and, if a longitudinal study, the frequency with which their participation will be requested), how the data will be utilized, and their rights as research participants, including their right to confidentiality. Based on the information provided by the researcher, potential respondents can then make an informed determination as to whether they are willing to participate in a given study (i.e., give their consent). In addition to the willingness to participate, it is fundamental that potential respondents have the competence to understand why the study is being conducted and what their rights and responsibilities are as respondents in order to participate.

Survey research among children and adolescents under the age of 18 requires parental or guardian consent before the researcher can even speak with the juvenile. Frequently parental consent must be given in writing through use of either an active or passive consent form. Also, some ethics boards may require, in addition to parental consent, the assent of juvenile participants. If assent is required of juvenile participants, they must also have the capacity to comprehend the purpose of the study and their rights and responsibilities as participants.

Respondents' Rights

Respondents' rights are paramount to any ethical survey research project and an integral part of informed consent. Researchers have an obligation to their subjects to minimize the possibility of risk. Granted, respondents participating in the bulk of survey research are not generally at a high risk of physical or psychological harm. However, some survey research topics are very sensitive in nature and may cause a considerable amount of discomfort for some respondents. In addition, researchers are ethically bound to report any child abuse that is suspected. Therefore, it is essential to minimize risk through a thorough advance disclosure of any possible harm or discomfort that may result from survey participation. If there are compelling scientific reasons that respondents must be kept "blind" to, or entirely deceived about, some of the aspects of a study before they give their consent, and while the study is being conducted (e.g., an experimental design would be compromised if the respondents knew it was being conducted), then it is the responsibility of the researcher to debrief the respondents about any deception they may have experienced, even if it can be argued that the deception was trivial in nature. It also is the responsibility of the researcher to "undo any harm" he or she may have caused associated with any deception or other withholding of information.

Furthermore, respondents are afforded additional rights that also minimize any risks associated with their participation, including the right that their responses will be kept confidential and the right to privacy. Confidentiality protects respondents' identities so that their participation in a given study cannot be determined and, likewise, ensures that their responses are not linked to them personally. The importance of maintaining confidentiality is closely related to minimizing risk. For example, survey data on a study of criminal behavior could be subject to subpoena, but if identifier data are never collected or are destroyed, the individuals and their responses cannot be identified. Thus, the data collected through the survey research process should not contain identifying information about individual respondents, and data records should be securely stored and destroyed as soon as is appropriate.

Survey research is intrusive in nature: Respondents are asked to reveal personal information about themselves, their behaviors, and their beliefs. Thus, researchers must consider respondents' right to privacy when administering surveys. The right to privacy does not suggest that personal questions should not be asked of respondents (as long as they are relevant to the study being conducted), but it protects respondents from disclosing such information if they choose not to respond. In other words, respondents' right to privacy is the freedom afforded survey respondents to control the personal information that is disclosed, under what circumstances they will do so, and with whom such information is shared. Subjects must be informed that their participation is voluntary, that their refusal to participate will not involve any penalty, and that they may skip any question they do not feel comfortable answering or discontinue their participation at any time.

In some instances, respondents are provided incentives for their participation in a given study; however, this does not negate their right to skip certain questions or end their involvement in the study at their will. Furthermore, incentives should never be used to "coerce" (even subtly) a respondent to participate in a study or to answer a specific question if he or she really does not want to do so. Ethical researchers do not force, coerce, "seduce," trick, or otherwise threaten potential subjects when attempting to gain cooperation or administer a survey.

Social Responsibility

Social responsibility is key to survey ethics. Research findings contribute to the larger body of knowledge for purposes of better understanding social behavior and improving the quality of life among members of society. Thus, researchers have an obligation of being forthright not only with research participants but also with society in general. Deception in research practices and principles can lead to public distrust and bring even the best research under public and professional scrutiny. Ethics boards ensure, in principle, that a given study meets ethical guidelines. However, there is little ethical oversight once data collection has begun. As such, regulation and oversight through audits of individual researchers' or research firms' ethics policies and practices regarding survey research could potentially decrease public deception and potential harm.

Researchers have an obligation to protect and respect not only the rights of research participants, but to society as a whole as well. Informed consent outlines the information necessary for respondents to make voluntary informed decisions about participating in a given study based on their understanding of what the study is about, how the data will be used, and what their rights and responsibilities are as participants. Research participants' rights are paramount when conducting survey research. If the study cannot be designed ethically, then it should not be conducted. The continuation of survey research as an invaluable tool for gathering information is contingent upon maintaining public confidence which can only be accomplished through upholding ethical practices and principles.

Lisa M. Gilman

See also Confidentiality; Consent Form; Debriefing; Deception; Ethical Principles; Incentives; Informed Consent; Institutional Review Board (IRB); Privacy; Voluntary Participation

Further Readings

Areen, J. (1991). Legal constraints on research on children. In B. Stanley & J. E. Sieber (Eds.), *The ethics of research on children and adolescents*. Newbury Park, CA: Sage.

Israel, M., & Hay, I. (2006). *Research ethics for social scientists*. Thousand Oaks, CA: Sage.

Sales, B. D., & Folkman, S. (Eds.). (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association.

Sieber, J. E. (1992). *Planning ethically responsible research: A guide for students and internal review boards*. Newbury Park, CA: Sage.

Survey Methodology

Survey Methodology is a peer-reviewed journal published twice yearly by Statistics Canada. The journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time-series analysis, seasonal adjustment, demographic methods, data integration, estimation and data analysis methods, and general survey systems development. Emphasis is placed on development and evaluation of specific methodologies as applied to data collection, data processing, estimation, or analysis.

The journal was established in 1975, as a Statistics Canada in-house journal intended primarily to "provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology." Its basic objectives and policy remained unchanged for about 10 years. During this period, however, the journal grew to the point that pressing demands and interests could not be met within the framework established at its inception.

In 1984 several major changes were introduced, which included broadening the scope of the editorial policy and expansion of the editorial board. A separate management board was also established. The editorial board now consists of the editor, a deputy editor, approximately 30 associate editors, and six assistant editors. M. P. Singh was the editor of the journal from the beginning until his death in 2005. The current editor is John Kovar. The associate editors are all leading researchers in survey methods and come from academic institutions, government statistical agencies, and private-sector firms around the world.

Since its December 1981 issue, one unique feature of the journal is that it is fully bilingual, with submissions being accepted in either English or French and all published papers being fully translated. The French name of the journal is *Techniques d'Enquête*.

From time to time the journal has included special sections containing a number of papers on a common theme. Topics of these special sections have included, among others, census coverage error, small area estimation, composite estimation in surveys, and longitudinal surveys and analyses. The June 2002 issue of the journal included a special section in honor of Leslie Kish.

In December 1999 and June 2000, *Survey Methodology* celebrated its 25th anniversary with two special issues containing invited papers from several prominent statisticians. The lead paper in the December issue (volume 25, number 2) by Richard Platek, founding chairman of the management board, gives an overview of the gradual evolution of the journal from its humble beginnings to a well-known international publication.

In 2000 the *Survey Methodology* journal, in cooperation with the American Statistical Association and Westat, established an annual invited paper series in honor of Joseph Waksberg. Each year a prominent survey researcher is chosen by a committee to write a paper reviewing the development and current state of a significant topic in the field of survey methodology. Winners of the Waksberg Award to date are Gad Nathan, Wayne Fuller, Tim Holt, Norman Bradburn, J. N. K. Rao, Alastair Scott, and Carl-Erik Särndal.

The journal is available on the Statistics Canada Web site, starting with the year 2000 issue up to the most current issue. Prior issues are available on request. Printed copies of the journal can be obtained by paid subscription.

John Kovar

See also Statistics Canada

Further Readings

Statistics Canada: http://www.statcan.ca

SURVEY SPONSOR

The sponsor of a survey is responsible for funding part or all of the sampling and data collection activities and typically has first or exclusive rights to the data. Often, survey sponsors do not have the in-house capacity (i.e., facilities or interviewing staff) to administer a survey and, as a result, contract these data collection activities to a third party. These third-party organizations focus on survey administration, such as selecting the sample, interviewing and distributing questionnaires, and data entry and analyses. Sometimes these third parties (i.e., survey research firms) are asked to provide expertise in questionnaire design and data analysis and asked to draft an analytic report. These contracted services do not negate the survey sponsor's exclusive (or first) rights to the survey data. Typically, a respondent is informed of both parties involved in conducting the survey—the data collection body and the survey sponsor. Identifying these different parties may impact response patterns.

Effects of Survey Sponsorship on Survey Accuracy

Several theories explain why survey sponsorship may affect response patterns, including the response rate, data quality, and responses. Government-sponsored research may produce higher response rates, because people think their responses are required by law and think that the government can compel a response to the survey, such as in population censuses. (In general, survey participation tends to be voluntary, but in some instances, survey participation is required by federal law in the United States and elsewhere.) These perceptions could impact responses and data quality. Perceptions that responses are required may lead to less-motivated respondents completing the survey, which may impact data quality, leading to more skipped questions or less-thoughtful responses. In addition, respondents may respond to these surveys differently when the government sponsor is identified or made salient because they fear that their responses may affect their access to government services. Alternatively, other people may be more likely to respond to government-sponsored surveys and provide more complete responses because they believe these survey findings are more likely to have a direct effect on them through policy changes.

Government and university-sponsored surveys may produce higher response rates merely because these institutions have higher prestige, are a moral authority, or are more "official" than other types of sponsors, such as private-sector firms. People may also be more willing to participate in government or university-sponsored surveys because they recognize that these types of surveys are not disguised sales calls. Furthermore, university-sponsored surveys may produce higher response rates, as well as more complete and thoughtful responses, because people believe they are contributing to the advancement of science and knowledge.

Private-sector companies and politicians, for example, share concerns about response bias and response quality, but they also have concerns around disclosing their sponsorship of a survey. These types of sponsors may be reluctant to disclose themselves as a sponsor to a sampled respondent—at least prior to the survey being completed—fearing that this information may bias survey responses, with respondents conveying their opinions about the sponsor rather than the survey questions. In addition, private-sector companies and politicians may be reluctant to identify their sponsorship out of fear that their competitors may learn about their proprietary market research.

Prior to the 1990s, several experiments were conducted related to the effects of the survey sponsor on response rates, data quality, and response bias. In general, the studies found that government or university sponsorship yielded higher response rates compared to sponsorship by a commercial organization. This body of research is not consistent about whether university sponsorship or government sponsorship produces a higher response rate. A well-cited British study by Christopher Scott suggested that government sponsorship yielded a slightly higher response rate over the London School of Economics or British Market Research Bureau (a commercial agency). Thomas A. Heberlein and Robert Baumgartner's meta-analysis in the 1970s found that government-sponsored research produced higher initial response rates, but controlling for topic saliency and other factors, sponsorship had no significant effect on the final response rate. John Goyder's test of the Heberlein and Baumgartner model of predicting response rates, using a more expansive database of research, found that government-sponsored surveys have significantly higher response rates.

Although many studies document differences in response rates depending on sponsorship, many researchers remain unconvinced that survey sponsor has a significant effect on response patterns. The advantage of noncommercial sponsorship (university, government) may diminish if the survey sponsor is identified as the survey organization conducting the survey (rather than a separate entity that is paying for the study). Commercial survey research firms (when no additional sponsor is identified) may produce slightly lower response rates than university-administered surveys, but these differences may not be statistically significant.

Although research tends to suggest an effect of sponsorship on response rates, the effects of sponsorship on data quality is more evenly split between finding no significant differences and finding that university sponsorship improved data quality by reducing item nonresponse. Several studies have compared survey responses under different sponsorship conditions to determine whether the responses differ, depending on the sponsor. This body of research suggests some small differences in responses to specific questions, but overall, responses to the survey tend to not differ, depending on sponsorship condition. In sum, sponsorship effects on response patterns may be more pronounced in participants' decisions whether to participate in a survey rather than in their responses to survey questions.

Shelley Boulianne

See also Bias; Missing Data; Response Rates

Further Readings

Goyder, J. C. (1982). Further evidence on factors affecting response rates to mailed questionnaires. *American Sociological Review*, *47*(4), 550–553.

Heberlein, T. A., & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, *43*(4), 447–462.

Scott, C. (1969). Research on mail surveys. *Journal of the Royal Statistical Society, 124 Series A, Part 2*, 143–191.

Systematic Error

Systematic errors result from bias in measurement or estimation strategies and are evident in the consistent over- or underestimation of key parameters. Good survey research methodology seeks to minimize systematic error through probability-based sample selection and the planning and execution of conscientious survey design. There are two key sources of systematic error in survey research: sample selection and response bias.

In the context of survey research, systematic errors may be best understood through a comparison of samples in which respondents are randomly selected (i.e., a probability sample) and samples in which respondents are selected because they are easily accessible (i.e., a convenience sample). Consider, for example, a research project in which the analysts wish to assess attitudes about a town's public library. One way to proceed might be to post students in front of the main entrance to the library and to have them survey individuals as they enter and leave the library. Another strategy would be to randomly assign households to participate in the survey, in such a way that every household in the town has a nonzero probability of participating, and to send students to the households to conduct the interview.

How would the results of the survey differ as a consequence of these differences in sample selection? One might reasonably expect that the first research design, in which the sample is composed entirely of current library patrons, yields a sample that uses the library more frequently and, as a consequence, has more favorable views about the library and its importance to the community than does the larger, more representative sample of the entire town. In this case, the selection bias inherent in the convenience sample of library patrons would lead the researchers to systematically overestimate the use of, and support for, the town's public library.

Although in the library example, systematic error resulted from the selection bias in the sampling mechanism, William G. Cochran, Frederick G. Mosteller, and John Tukey state that systematic errors more often result from bias in measurement. In the context of the physical sciences, it is easy to imagine a scale that consistently overestimates the weight of whatever it measures by five units. There are analogies to this scale in survey research: In election surveys, for example, researchers often want to identify the issues or problems most important to voters and assess their effect on voting decisions. Typically, survey items about the most important issue are asked as the open-ended question, *What do you think is the most important problem facing this country today?* An alternative format uses a closed-ended question, in which respondents are presented with a list of issues and asked how important they perceive each issue to be. The proportion of the sample that reports that they consider an issue to be "very important" in the closed-ended format is typically much larger than the proportion of the sample who identify the issue when asked the open-ended question. Social desirability, priming, and frame effects also shift responses in predictable ways and complicate the measurement of attitudes and opinions. Similarly, question structure and order effects can generate spurious patterns in survey responses that undermine the ability to evaluate public opinion.

Good survey research methodology strives to minimize both the selection bias and response bias that induce systematic error.

Systematic Error and Bias

The concept of systematic error is closely related to the more general statistical understanding of bias. To illustrate this relationship more concretely, suppose one is interested in estimating the population mean of the distribution of opinion on an issue that can be measured in a continuous way (e.g., the amount of funding to be devoted to a particular policy). After drawing a random sample of respondents and soliciting their opinion, a researcher estimates the sample mean and its variance. Under what conditions can the population mean be inferred from the sample mean, when considering the distribution of opinion on the issue the researcher cares about? The sample mean will be a misleading representation of the population mean if systematic error is incorporated in the sample selection mechanism—if the sample over- or underrepresents a subset of the population—or if the questions are constructed in a way that shifts responses away from the respondents' true opinions. Under these conditions, even if the researchers repeated the exercise of sampling and interviewing an infinite number of times, the sample means will never equal the population mean in expectation. Thus, as this example illustrates, bias can be attributed to systematic error in the measurement or estimation of key parameters.

Karen Long Jusko

See also Bias; Convenience Sampling; Probability Sample; Random Error; Response Bias; Self-Selection Bias

Further Readings

Cochran, W. G., Mosteller, F., & Tukey, J. W. (1954). Principles of sampling. *Journal of the American Statistical Association*, *49*(265), 13–35.

Green, D. P., & Citrin, J. (1994). Measurement error and the structure of attitudes: Are positive and negative judgments opposites? *American Journal of Political Science*, *38*(1), 256–281.

Schuman, H., Ludwig, J., & Krosnick, J. A. (1986). The perceived threat of nuclear war, salience, and open questions. *Public Opinion Quarterly*, *50*, 519–536.

Systematic Sampling

Systematic sampling is a random method of sampling that applies a constant interval to choosing a sample of elements from the sampling frame. It is in common use in part because little training is needed to select one. Suppose a sample of size n is desired from a population of size $N = nk$. Systematic sampling uses a random number r between 1 and k to determine the first selection. The remaining selections are obtained by taking every kth listing thereafter from the ordered list to yield the sample of size n. Because it designates the number of records to skip to get to the next selection, k is referred to as the *skip interval*.

Systematic sampling is particularly desirable when on-site staff, rather than a data collection vendor, select the sample. For example, one-page sampling instructions can be developed for a survey of mine employees that explain how the mine operator is to order the employees and then make the first and subsequent sample selections using a pre-supplied random starting point. Systematic sampling is also useful in sampling on a flow basis, such as sampling clients entering a facility to obtain services such as emergency food assistance or health care. In this situation, the facility manager is approached to determine how many clients might visit in the specified time period. This population estimate $\hat{N}$ is used to determine the skip interval k based upon the desired sample n of clients from the facility. At the end of specified time period, the actual population N of clients is recorded.

Frequently, the population size N, together with the desired sample size n, results in a skip interval k that is a real number as opposed to an integer value. In this case, the simplest solution is to round k down to create a skip interval that is an integer. This approach results in a variable sample size (n or $n + 1$), but it is preferable for use by nontechnical staff. Alternately, the list can be regarded as circular, with a random number between 1 and N selected and then a systematic sample of size n selected using the integer value of k obtained by rounding down. An easily programmable option is to select a real number between 1 and k as the random starting point and then continue adding the real number k to the random starting point to get n real numbers, which are rounded down to integers to determine the records selected for the sample.

Systematic sampling can be viewed as a form of implicit stratification. Conceptually, the frame is split into n zones each of size k and one selection made from each zone. When the frame is sorted by key analysis domains prior to selection, this implicit stratification results in a sample that is close to the results of stratification with proportional allocation. Care must be taken in ordering the sample and in using a preordered frame when systematic sampling is planned. A frame sorted by age, for instance, could produce samples that skew young or skew old, depending on the random start. The worst case scenario is a list that has an underlying periodicity to the observations, and this periodicity corresponds to the skip interval.

Unlike stratified sampling, systematic sampling is not independently executed across the zones or implicit strata. The initial selection from the first zone determines the selections made from the remaining $n - 1$ zones. Hence, systematic sampling can also be viewed as a form of cluster sampling, where the random start identifies a cluster of size n from the total of k possible clusters. The precision of survey estimates can be improved by sorting the frame prior to sample selection to get the benefits of implicit stratification and to create systematic clusters whose units are representative of the entire population of units.

Direct calculation of the variance requires that the systematic sampling process be replicated, so that it can be treated as a clustered design where each replicate is a cluster. Without replication, the variances of estimates derived from systematic samples are not estimable. Analysts typically approximate the variance by making assumptions about the ordered frame from which the systematic sample was selected. A common practice is to treat the systematic sample as if it were a simple random sample. Random ordering of the frame prior to systematic selection would have produced a simple random sample, for instance. This simple random sample assumption may over- or underestimate the variance of survey estimates depending on the intracluster correlation induced by the frame ordering. This approach is likely to produce conservative estimates (overestimates) of the variance for well-designed systematic samples with judiciously chosen sort variables that induce beneficial implicit stratification.

Sometimes, it may be appropriate to regard the systematically sampled units as being derived from a stratified sample with two units per strata. The systematic ordering of the population is used to define these pseudo-strata. This technique is most commonly used when systematic sampling has been used to select a probability-proportional-to-size sample, such as primary sampling units for an area household survey. Here, the skip interval and random starting point for the systematic sample are based upon the units' size measures, which are accumulated across units to yield the aggregated total M.

Brenda G. Cox

See also Cluster Sample; n; N; Probability Proportional to Size (PPS) Sampling; Proportional Allocation to Strata; Sampling Frame; Simple Random Sample; Stratified Sampling

Further Readings

Buckland, W. R. (1951). A review of the literature on systematic sampling. *Journal of the Royal Statistical Society B 13*, 208–215.

Iachan, I. (1982). Systematic sampling: A critical review. *International Statistical Review*, *50*, 293–303.

T

TAILORING

Tailoring is a term that is used in different ways in social behavioral research. It often is used to describe health behavior and health education messages that have been crafted to appeal to and influence the individual, with the intent of modifying the individual's attitudes and behavior such that he or she engages in healthier endeavors. A message that is tailored is a personalized message that attempts to speak to the individual *as an individual* and not as a member of any group or stratum. Additionally, a tailored message attempts to create a customized, personally meaningful communication—based on knowledge (data and information) about individuals. A more formal definition comes from M. W. Kreuter and C. S. Skinner and indicates that a tailored message is any combination of information or change strategies intended to reach one specific person, based on characteristics that are unique to that person, related to the outcome of interest, and have been derived from an individual assessment. This definition highlights the two features of a tailored health promotion intervention that distinguish it from other commonly used approaches: (a) Its collection of messages or strategies is intended for a particular person rather than a group of people, and (b) these messages or strategies are based on individual-level factors that are related to the behavioral outcome of interest. Tailoring techniques are utilized also by marketing practitioners but with the marketing goal of selling consumer goods and services.

Tailored messages can be based on age, sex, educational attainment, and other sociodemographics, as well as cognitive, motivational, and behavioral attributes. These messages usually refer to interventions created specifically for individuals with characteristics unique to them, and they are usually based on data collected from them as well.

Tailoring Versus Targeting

A closely related term that is often used interchangeably with tailoring is *targeting*. Although controversy and debate over definitions continue, the two strategies are clearly distinguishable. Targeting involves identifying a specific population, such as a group with a high rate of smoking prevalence, and designing a program, intervention, or marketing campaign that includes specific messages or materials intended for that group of smokers. For example, a targeted health message might attempt to influence these people to stop smoking. So, this kind of message can be thought of as focusing on the group attribute of smoking behavior, at least in terms of how the intervention could be focused. Tailoring, on the other hand, could be an approach to health messaging that could also be utilized, but one that focuses on more individual attributes. If health organizations were able to discern people's ages, their sex, or educational attainment, then health messaging might be made more "personalized" and thus more "meaningful" to the individual by creating more focused messages. The point where a group attribute becomes an individual attribute is when the message begins to be segmented or subdivided by individual attributes. While the point where that occurs still may be debated, targeted interventions can be thought of as strategies based on group-level variables and tailored interventions as strategies and messages that are customized based on individual-level variables.

Theoretical Approach

Tailored communications are not theory-specific, in that many different kinds of theory might be used to anchor an approach to attitudinal or behavioral modification. The important point is that a theoretical approach is utilized, as a systematic means to achieve a desired end, and different theories are best used as a basis for tailoring in relation to different behaviors and audiences. The parameters of a message—the quality, appearance, literacy level, and packaging of tailored communications—can vary widely. In the ideal, tailored materials eliminate all superfluous information and just transmit the right message to the right individual, through the right channel at the right time. Of course, reality is far from that ideal, as few behaviors have been subject to enough rigorous theory-based research. It is the utilization of theory that guides a reasoned and grounded approach that clearly delineates what is superfluous and what part of the message contains the "active ingredients," that is, the parts that get people to think, act, buy, cooperate with a survey request, and so forth. Moreover, tailored materials can be designed well or poorly. So, just because tailoring is used as a messaging technique does not mean that other elements of creative design become meaningless. In fact, those other elements are likely to become more important. Tailoring holds great promise and great challenges as well. For example, tailoring population-level interventions so that they also are personalized at the individual level could have a wide-scale impact on public health outcomes.

Application to Public Health

Using tailored health messages is an emerging strategic area for health communications, which attempts to maximize overall effects by striking an effective balance between personalization and cost. Multiple messages can be delivered over time, and because communication does not require one-on-one interpersonal interaction, costs may be greatly reduced. Information on the individual may be gathered through a survey or through a brief, personal interview with a professional. The results of these questions are entered into a computer, which draws from a "library" of possible messages to create materials that directly address the individual's needs, interests, or concerns. Once a program has been developed for a certain health issue, it can be used to produce tailored materials with the potential to reach large populations. Thus, tailored communications are an ideal tool to address a variety of public health issues.

Although the field of tailoring is still in its infancy, empirical research shows that tailored print materials are more effective than nontailored ones in helping people change health-related behaviors such as smoking, diet, and physical activity. According to Skinner and colleagues, who published the first scientific review of tailored research, not only are tailored (print) communications more effective than nontailored ones, but also they can be an important adjunct to other intervention components.

There are many exciting future directions and huge potential for tailorized approaches. Constantly evolving personal technology, such as Internet Web sites, email, instant messaging, personal desktop assistants, cell phones, and computerized kiosks, to name just a few, will make it easier to administer behavioral interventions on a wide scale, changing the landscape of public health in a dramatic fashion.

Special Application Within Survey Research

Survey researchers have developed an innovative method that utilizes tailoring to improve response rates. Following from leverage-saliency theory, an interviewer will be most likely to gain cooperation from a reluctant respondent if the interviewer "tailors" his or her introductory spiel to the specific concerns that the respondent expresses (or appears to be expressing). Linked to this, refusal avoidance training (also called *interviewer refusal aversion training*) curricula are ones that build on an understanding of this interviewer–respondent interaction to help interviewers deal effectively with respondent reluctance and avoid refusals by tailoring persuasive strategies to answer the callers' specific questions and keeping them talking. What is theorized to occur with inexperienced or unskilled interviewers is that they either press the respondent to make a decision too quickly, or they do not give an effective response to specific respondent concerns or questions, thus, leading to a refusal. Because experienced and skilled interviewers are more adept at tailoring their answers to the specific household they are calling, while maintaining continued conversation with the caller, they are more successful in getting respondents to complete the survey. Robert M. Groves and Mick P. Cooper have described a training protocol that can be grounded in this theory. In their model, the following components are necessary for tailoring to be successful: (a) The interviewer must have a repertoire of techniques, strategies, phrases, and so forth, related to the particular survey request; (b) the interviewer must be adept at reading the verbal and nonverbal cues from the respondent; (c) the interviewer must be able to apply the appropriate strategy according to the cues received from the respondent; and (d) the interaction between the interviewer and respondent must be long enough so that tailoring can be applied.

Joseph E. Bauer

See also Interviewer Training; Leverage-Saliency Theory; Refusal Avoidance Training (RAT)

Further Readings

Glanz, K., Rimer, B. K., & Lewis, M. (Eds.). (2002). *Health behavior and health education: theory, research, and practice* (3rd ed.). New York: Wiley.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*, 475–495.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Kreuter, M., Farrell, D., Olevitch, L., & Brennan, L. (2000). *Tailored health messages: Customizing communication with computer technology*. Mahwah, NJ: Lawrence Erlbaum.

Kreuter, M. W., & Skinner, C. S. (2000). Tailoring, what's in a name? *Health Education Research*, *15*, 1–4.

Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin*, *133*(4), 673–693.

Skinner, C. S., Campbell, M. K., Rimer, B. K., Curry, S., & Prochaska, J. O. (1999). How effective is tailored print communications? *Annals of Behavioral Medicine*, *21*(4), 290–298.

Target Population

The target population for a survey is the entire set of units for which the survey data are to be used to make inferences. Thus, the target population defines those units for which the findings of the survey are meant to generalize. Establishing study objectives is the first step in designing a survey. Defining the target population should be the second step.

Target populations must be specifically defined, as the definition determines whether sampled cases are eligible or ineligible for the survey. The geographic and temporal characteristics of the target population need to be delineated, as well as types of units being included. In some instances, the target population is restricted to exclude population members that are difficult or impossible to interview. For instance, area household surveys tend to define their target populations in terms of the civilian, noninstitutionalized population of the United States. Any exclusion made to the target population must also be reflected in the inferences made using the data and the associated presentations of findings. Undercoverage of the target population occurs when some population units are not linked or associated with the sampling frame and hence have no chance of inclusion in the survey. The subset of the target population that has a chance of survey inclusion because of their membership in, or linkage to, the sampling frame is referred to as the *survey population*. Traditional telephone surveys have a survey population of households with landline telephone service but typically are used to make inferences to target populations of all households, regardless of telephone service. In some instances, surveys have more than one target population because analysis is planned at multiple levels. For instance, a health care practices survey might have a target population defined in terms of households, another defined in terms of adults, and another defined in terms of children. Such a survey might sample households, collect household-level data, and then sample an adult and a child for interviews at those levels.

For business surveys, the target population definition must also specify the level of the business that comprises the units of the target population. Surveys of business finances typically define their target populations in terms of the enterprise—the organizational level that has ownership and ultimate responsibility for making decisions for the entire business. Labor force surveys define their target populations in terms of the establishment, that is, at the organizational level, where employment activities are conducted at or from a particular geographic location.

Brenda G. Cox

See also Establishment Survey; External Validity; Sampling Frame; Undercoverage

Further Readings

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

Nijhowne, S. (1995). Defining and classifying statistical units. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business survey methods* (pp. 49–64). New York: Wiley.

Taylor Series Linearization (TSL)

The Taylor series linearization (TSL) method is used with variance estimation for statistics that are vastly more complex than mere additions of sample values.

Two factors that complicate variance estimation are complex sample design features and the nonlinearity of many common statistical estimators from complex sample surveys. Complex design features include stratification, clustering, multi-stage sampling, unequal probability sampling, and without replacement sampling. Nonlinear statistical estimators for complex sample surveys include means, proportions, and regression coefficients. For example, consider the estimator of a subgroup total, $\hat{y}_d = \sum_i w_i d_i y_i$, where w_i is the sampling weight, y_i is the observed value, and d_i is a zero/one subgroup membership indicator for the ith sampling unit. This is a linear estimator because the estimate is a linear combination of the observed values y_i and d_i. On the other hand, the domain mean, $\hat{\bar{y}}_d = \sum_i w_i d_i y_i / \sum_i w_i d_i$, is a nonlinear estimator as it is the ratio of two random variables and is not a linear combination of the observed data.

Unbiased variance estimation formulae for linear estimators are available for most complex sample designs. However, for nonlinear estimators, unbiased variance estimation formulae are often not available, and approximate methods must be used. The most common approximate methods are replication methods, such as the *jackknife method* or *balanced repeated replication*, and the TSL method.

The TSL method uses the linear terms of a Taylor series expansion to approximate the estimator by a linear function of the observed data. The variance estimation formulae for a linear estimator corresponding to the specific sampling design can then be applied to the linear approximation. This generally leads to a statistical consistent estimator of the variance of a nonlinear estimator.

To illustrate the TSL method, let $\hat{\boldsymbol{\theta}} = F(\hat{y}, \hat{x})$ be an estimate of the parameter $\boldsymbol{\theta}$ where $\hat{y}$ and $\hat{x}$ are two linear sample statistics. For example, $\hat{\boldsymbol{\theta}} = \hat{\bar{y}}_d$. Also define μ_y and μ_x to be the expected values of $\hat{y}$ and $\hat{x}$, respectively. $\hat{\boldsymbol{\theta}}$ can be expanded in a Taylor series expansion about μ_y and μ_x so that

$$\hat{\boldsymbol{\theta}} = F(\mu_y,\mu_x) + (\partial F_y)(\hat{y} - \mu_y) + (\partial F_x)(\hat{x} - \mu_x) \\ + \text{higher-order terms},$$

where ∂F_y and ∂F_x are the first-order partial derivatives of F with respect to $\hat{y}$ and $\hat{x}$ evaluated at their respective expectations, μ_y and μ_x. If the higher-order terms are negligible, then variance of $\hat{\boldsymbol{\theta}}$ can be approximated by

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\theta}}) &\cong E[\hat{\boldsymbol{\theta}} - F(\mu_y,\mu_x)]^2 \\ &= (\partial F_y)^2 E(\hat{y} - \mu_y)^2 + (\partial F_x)^2 E(\hat{x} - \mu_x)^2 \\ &\quad + 2(\partial F_y)(\partial F_x)E[(\hat{y} - \mu_y)(\hat{x} - \mu_x)] \\ &= (\partial F_y)^2 \mathrm{Var}(\hat{y}) + (\partial F_x)^2 \mathrm{Var}(\hat{x}) \\ &\quad + 2(\partial F_y)(\partial F_x)\mathrm{Cov}(\hat{y},\hat{x}). \end{aligned}$$

This approximation can easily be extended to functions of more than two linear sample statistics. An approximate estimate of the variance of $\hat{\boldsymbol{\theta}}$ is then obtained by substituting sample-based estimates of μ_y, μ_x, $\mathrm{Var}(\hat{y})$ and $\mathrm{Var}(\hat{x})$ in the previous formula.

An equivalent computational procedure is formed by recognizing that the variable portion of the Taylor series approximation is $\hat{z} = (\partial F_y)\hat{y} + (\partial F_x)\hat{x}$ so that

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) \cong \mathrm{Var}[(\partial F_y)\hat{y} + (\partial F_x)\hat{x}] = \mathrm{Var}(\hat{z}).$$

Because $\hat{y}$ and $\hat{x}$ are linear estimators, the Taylor series variance approximation can be computed using the linearized values $z_i = w_i[(\partial F_y)y_i + (\partial F_x)x_i]$ so that

$\hat{z} = \sum_i z_i$. As before, substituting sample-based estimates of μ_y and μ_x, namely, $\hat{y}$ and $\hat{x}$, in the formula for z_i and then using the variance formula of a linear estimator for the sample design in question to estimate the variance of $\hat{z}$ yields an approximate estimate of the variance of $\hat{\theta}$. This reduces the problem of estimating the variance of a nonlinear statistics to that of estimating the variance of the sum of the linearized values. As an example, the linearized values for the mean $\hat{\bar{y}}_d$ are $z_i = w_i d_i (y_i - \hat{\bar{y}}_d) / \sum_i w_i d_i$.

This illustration of the TSL method was for an estimator that is an explicit function of the observed data such as a mean, proportion, or linear regression coefficient. There are extensions of the TSL method to estimators that are implicitly defined through estimating equations, such as the regression coefficients of logistic, log-link, or Cox proportional hazards models.

Rick L. Williams

See also Balanced Repeated Replication (BRR); Clustering; Complex Sample Surveys; Jackknife Variance Estimation; Multi-Stage Sample; Probability of Selection; Replicate Methods for Variance Estimation; Sampling Without Replacement; Stratified Sampling; SUDAAN

Further Readings

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279–292.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Wolter, K. M. (2007). *Introduction to variance estimation*. New York: Springer.

Woodruff, D. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, *66*, 411–414.

Technology-Based Training

Technology-based training uses computer-based tools to enhance the training process, typically by involving trainees actively rather than passively. For survey research, technology-based training is usually used to build and sharpen interviewer skills, particularly skills required for successfully interacting with survey respondents. Traditionally, interaction-skills training relied on peer-to-peer role playing or passive learning through videos. Technology-based training, in contrast, facilitates self-directed learning with rich media sources and abundant learner-assessment and remediation options. The benefits of technology-based training for survey research include (a) reduced learning time; (b) reduced or eliminated travel time and expense; (c) improved consistency by capturing and replicating best practices and expert knowledge; (d) increased availability of training (with just-in-time access on personal computers in any location); (e) enhanced productivity by decreasing on-the-job error rates and reducing reliance on on-the-job learning; and (f) increased ability to adapt to interviewers' knowledge, experience, learning style, and motivation.

Some technology-based training programs are based on a programmed learning model: Information is delivered through multimedia (text, graphics, video, and narration), and trainees' understanding of the material is tested through multiple-choice questions and other basic evaluation methods. This approach usually involves breaking large blocks of training content into discrete modules that can be searched electronically and then studied by the trainee in a short time. Breaking down a skill into many component parts allows users to gain competency quickly. Learning objectives also help users who need to brush up on a skill when they are back on the job. Trainees can skim through the list of topics in a course module and find the answer they need immediately without wading through many pages of text.

In contrast, an experiential learning model emphasizes trainees' actual experiences as the starting point of the training process. The mental processes used to analyze these experiences are also stressed. New technologies also allow trainees' skills to be developed and practiced in realistic settings with realistic spoken interaction. For example, *virtual reality* is a realistic, three-dimensional, interactive simulation of the trainees' work environment, and *natural language processing* allows trainees to speak into a microphone and have the computer application recognize their words, interpret them in context, determine their meaning, and generate an appropriate response. Such approaches increase the time trainees spend acquiring and practicing critical skills, increase active learning (which allows trainees to retain skills better than does passive learning), improve the realism of practice sessions, and facilitate intelligent tutoring. A training session in virtual reality typically includes (a) instruction on the scope of the task; (b) a definition of the goals and objectives; (c) a representation of an environment through visual, auditory, and at times kinesthetic information; (d) control systems to determine how the learner interacts with the simulation; (e) embedded instruction about content and process; and (f) coaching assistance.

Flexibility and responsiveness are critical for developing effective interaction skills and for performing well under difficult conditions, such as in a limited time or with limited information. To acquire flexible and effective skills at gaining respondents' cooperation, new and experienced interviewers require a learning environment that realistically simulates the environment they face in an interviewing situation. The consistency that is gained by repetitive practice in virtual and constructive learning environments leads directly to effective decisions in the production environment. Practice also leads to increased confidence before the first on-the-job experience, minimizing the amount of on-the-job learning necessary.

In the survey world, on-the-job-learning can translate into numerous unsuccessful interview attempts by a new interviewer at the start of a study, leading to lower response rates, lower-quality data, delayed schedules, and increased costs. This is exactly the scenario that *virtual training environments* can be most effective at preventing. Generally, interviewing skills are taught through a combination of lecture, mock practice sessions with other interviewer trainees, and audiotapes of real or mock exchanges between interviewers and survey respondents. Virtual environments, however, facilitate skill building at a higher level, by providing a simulated but realistic environment in which interviewers can practice and hone their skills.

Use of virtual environments for skill building also has disadvantages. Most importantly, current technology does not produce fully realistic conversational partners. Advances in natural language dialogue features and behavior models will add tremendously to the realism as this technology evolves. In addition, technology for training works best when the technology is used to enhance, rather than replace, well-prepared training materials. If the training program itself is badly conceived, sophisticated technology will not rescue it. Conversely, inappropriate use of technology can make a good training program less effective.

Michael Link

See also Interviewer Training

Further Readings

Link, M., Armsby, P., Hubal, R., & Guinn, C. (2006). Accessibility and acceptance of responsive virtual human technology as a telephone interviewer training tool. *Computers in Human Behavior*, *22*, 412–426.

TELEMARKETING

Organizations engage in various forms of direct marketing to sell their products and services, solicit money, or nurture client relationships. Telemarketing is one such direct marketing technique that can target either individual consumers or other organizations. Groups that engage in telemarketing include nonprofit organizations, institutions, and political interest groups. A telemarketing call could come from a charity soliciting a small donation, an alumni group from a university, or a political party seeking support for a candidate. However, the prototypical use of telemarketing is for purposes of private enterprise. For businesses, telemarketing is a common form of direct selling, which originated with the advent of the telecommunications industry. Though it had been used with varying degrees of success for decades, it was not until the 1970s that the telephone became a standard tool of mass marketing. *Cold calling*, the unsolicited calling of large groups of people, is just one form of profit-seeking telemarketing. It should be recognized that although many firms do engage in cold calling, the two primary technological advances that popularized telemarketing practice are not closely tied to this practice.

By the late 1970s, AT&T widely disseminated toll-free numbers, which allowed for customers to call *into* a business to gather more product information, receive specific services, and place long-distance orders free of a telephone line usage fee. The second technology that can be credited with the rise of telemarketing is the electronic management of database information. As computers increased in capability and declined in price, businesses capitalized on the ability to marry customer information and phone numbers, with their own product catalogs. Consequently, while cold calling is a "shot in the dark" for businesses, the use of the phone can take on more strategically focused forms. For example, telemarketers can manage customer accounts or screen the person on the phone for characteristics that would qualify him or her for different sales techniques. Product information, such as marketable innovations, can be relayed to past customers. Further, market surveys are completed over the telephone. For instance, at the point of sale a retail consumer can be provided with a toll-free number and a unique identification code that he or she can use to access an automated phone survey system using interactive voice response software.

Telemarketing is not without its faults and criticisms, and it is often perceived to carry a worse reputation with the public than mail- or Internet-based contact methods. Because of its personal yet faceless nature, telemarketing has been used fraudulently and unethically. One example of this is with unscrupulous magazine selling, which can trap consumers into sales agreements that stretch on for years at above-market prices. This happens, in part, because laws in a number of states require nothing more than verbal consent for a consumer to be legally bound to paying for a magazine subscription. Another pitfall in magazine sales, and otherwise, is the intentional incomplete disclosure of sales terms and prices over the telephone while the salesperson continues to seek a commitment from the possible buyer. However, it is the claimed invasion of privacy into the household that most frequently draws the ire of the phone-owning masses. In response to this, the Federal Trade Commission manages the National Do Not Call Registry, which allows people to submit their phone number for exclusion from most types of unsolicited sales calls.

Many believe that survey research response rates have been seriously harmed by the effects of telemarketing. As public annoyance with unsolicited telemarketing has grown, more and more citizens are refusing legitimate telephone survey contacts, in part because they do not differentiate the two types of call. Survey research professional organizations, such as CMOR(Council for Marketing and Opinion Research) and CASRO (Council of American Survey Research Organizations), strive to educate the public so that legitimate surveys are not confused with telemarketing. These efforts have not yet been very successful.

Matthew Beverlin

See also Council for Marketing and Opinion Research (CMOR); Council of American Survey Research Organizations (CASRO); Do-Not-Call Registries; FCC Regulations; FRUGing; FTC Regulations; Interactive Voice Response (IVR); SUGing

Further Readings

Federal Trade Commission: http://www.ftc.gov
National Do Not Call Registry: https://www.donotcall.gov/default.aspx
National Fraud Information Center: http://www.fraud.org/telemarketing/teleinfo.htm
Osborne, G. S. (1984). *Electronic direct marketing*. Englewood Cliffs, NJ: Prentice Hall.

Telephone Consumer Protection Act of 1991

In 1991, the U.S. Congress, in response to concerns and complaints about the increasing number of unsolicited telemarketing calls to consumers, passed the Telephone Consumer Protection Act (TCPA). The TCPA (Public Law 102–243) updated the Communications Act of 1934 and is the primary law governing telemarketing. The TCPA mandated that the Federal Communications Commission (FCC) amend its rules and regulations to implement methods for protecting the privacy rights of citizens by restricting the use of the telephone network for unsolicited advertising as stated in the TCPA. Although the TCPA was specifically directed at telemarketing activities and abuses, some of its prohibitions and some of the subsequent FCC "Rules and Regulations Implementing the Telephone Consumer Protection Act of 1991" are not specific to telemarketing. Rather they are aimed at protecting individual privacy in general and, as such, can impact telephone survey research.

TCPA

The TCPA specifically restricted the use of automatic telephone dialing systems, prerecorded voice messages and unsolicited fax advertisements:

> It shall be unlawful for any person within the United States to (A) make any call (other than a call made for emergency purposes or made with the prior express consent of the called party) using any

automatic telephone dialing system or an artificial or prerecorded voice to any emergency telephone line (including any "911"), to the telephone line of any guest room or patient room of a hospital, health care facility, elderly home, or similar establishment; or to any telephone number assigned to a paging service, cellular telephone service, specialized mobile radio service, or other radio common carrier service, or any service for which the called party is charged for the call; (B) to initiate any telephone call to any residential telephone line using an artificial or prerecorded voice to deliver a message without the prior express consent of the called party, unless the call is initiated for emergency purposes or is exempted by rule or order by the Commission under paragraph (2)(B); (C) to use any telephone facsimile machine, computer, or other device to send an unsolicited advertisement to a telephone facsimile machine; or (D) to use an automatic telephone dialing system in such a way that two or more telephone lines of a multi-line business are engaged simultaneously.

The TCPA and the FCC granted individuals and states the right to sue for damages in state court for individual violations of its rules and regulations and granted states the right to take civil action in federal district court against telemarketers "who engage in a pattern or practice of violation." The TCPA also allowed the FCC to define exemptions to some of the rules. Finally, the TCPA required the FCC to develop regulations to implement methods and procedures for protecting the "residential telephone subscribers' privacy rights to avoid receiving telephone solicitations to which they object."

FCC Reports and Orders

On September 16, 1992, the FCC adopted its first report and order (FCC 92-443), which established rules and regulations implementing the Telephone Consumer Protection Act of 1991. With the exception of the rule prohibiting the use of automatic telephone dialing systems, the FCC rules specifically refer to "telephone solicitation" and "unsolicited advertisement," more commonly known as telemarketing. Some exemptions were made to these rules for "tax-exempt nonprofit organizations" and "established business relationships." Of note, the Federal Trade Commission (FTC) has defined research as "informational" and not "telemarketing" and as such cannot be granted, nor does it need, exemption status.

In this first order the FCC opted for "the most effective and efficient" solution for complying with the TCPA requirement that they define procedures allowing subscribers to avoid unwanted solicitations—company-specific *do-not-call lists*. This first order also required telephone solicitors to have a written policy for maintaining a do-not-call list for the purpose of eliminating from their future solicitations any person who requested they not be called again by this organization. Do-not-call lists were to be maintained indefinitely. In 1995, responding to complaints from the telemarketing industry, the FCC modified its do-not-call rules. The new order required that do-not-call requests had to be honored for 10 years instead of indefinitely as originally ruled. In June of 2003, following the establishment of the National Do Not Call Registry, the FCC adopted another report and order (FCC 03-153), which established call abandonment rules, caller ID rules, and national do-not-call rules.

Automatic Telephone Dialing Systems

Initially the TCPA rules regulating the use of automatic telephone dialing systems had little or no impact on random-digit dialing telephone surveys. Most emergency numbers, hospital numbers, and wireless numbers were either in dedicated prefixes or blocks (not in a list-assisted frame) or removed as out-of-scope business numbers during sample generation. All this changed with the implementation of local number portability. In 1996 the U.S. Congress amended the Telecommunications Act of 1934 to establish a national framework that would promote competition for telephone service. One of the major barriers to competition had been the inability of customers to switch from one telephone service provider to another while retaining the same phone number. Local number portability is the ability of users of telecommunications services to keep their existing telephone number when changing from one service provider to another.

Local number portability was implemented in stages. By 2004 full portability had been implemented, allowing porting between all types of service: wireline to wireline, wireless to wireless, wireless to wireline, and wireline to wireless. At this point telephone survey research was affected because a wireline or landline telephone number appearing in a telephone directory or in a random-digit dialing sample might have been ported to a wireless phone, and dialing such numbers using automated telephone equipment in the United States would violate TCPA regulations. Given the significant financial penalty for such a violation, telemarketers and the research industry lobbied the FCC to provide a method for identifying these numbers. *NeuStar*, the designated administrator of the national database of ported numbers, and the FCC agreed to license wireless porting information for an annual fee. Licensees have access to two files of ported numbers, which are updated daily: a file of wireless-to-wireline telephone numbers and a file of wireline-to-wireless. These files are used by sample providers, research firms, and telemarketing firms to comply with TCPA regulations.

This TCPA prohibition has impacted telephone survey research in another emerging arena: cell phone sampling. As the number of households that have only a cell phone increases (approximately 20% by the end of 2007, with another 10% of households using their cell phones almost exclusively to receive calls even though they also have a landline), researchers will need to develop sampling methodologies that will include cell phone only and "cell phone mostly" households. The requirement that cell phone numbers must be hand-dialed in the United States will clearly add to survey costs and complexity.

Do-Not-Call Legislation

The TCPA had required the FCC to "compare and evaluate alternative methods and procedures (including the use of electronic databases, telephone network technologies, special directory markings, industry-based or company-specific 'do-not-call' systems, and any other alternatives, individually or in combination) for their effectiveness in protecting such privacy rights, and in terms of their cost and other advantages and disadvantages." Originally the FCC had opted for company-specific do-not-call lists. In the absence of a national database, many states enacted their own do-not-call legislation and lists. Some states also implemented special directory markings. Consumers still were not satisfied.

Responding to ongoing consumer complaints, the Do-Not-Call Implementation Act of 2003 (Public Law No. 108-10) was signed into law on March 11, 2003. This law established the FTC's National Do Not Call (DNC) Registry in order to facilitate compliance with the Telephone Consumer Protection Act of 1991. The FTC is responsible for the enforcement of the Telemarketing Sales Rule, which was mandated by Congress through the Telemarketing and Consumer Fraud and Abuse Prevention Act of 1994 and had been in effect since December 31, 1995. The Telemarketing Sales Rule was enacted to protect consumers from deceptive and abusive telemarketing practices. This legislation gives the FTC and state attorneys general law-enforcement tools to combat telemarketing fraud and levy penalties for violations and abuses.

The FTC's National Do Not Call Registry went into effect on October 1, 2003. Because the FCC and FTC regulate different components of the telecommunications industry (interstate and intrastate respectively), the FCC redefined its do-not-call rules in its 2003 report and order to implement the Do-Not-Call Implementation Act of 2003. FCC rules related to company-specific do-not-call lists remained in place. The National Do Not Call Registry is, as the name implies, national in scope and applies only to telemarketing calls. Telemarketers are required to download and "scrub" their lists against the national registry at least every 31 days. Exemptions have been made in the legislation for political organizations, charities, organizations with an established business relationship with the consumer or prior consent to call. It is important to remember that survey research is not formally exempted; it is simply not covered by this legislation. In other words telephone survey research is implicitly, not explicitly, exempt.

Access to the National Do Not Call Registry is limited to a "seller" of goods and services, a "telemarketer," a "service provider" (defined as a person or business that provides assistance to sellers or telemarketers) or an "exempt organization" (defined in the previous paragraph). In order to access the registry, an organization must pay the appropriate fee and agree to the certification requirements to receive a subscription account number with the National Do Not Call Registry. Certification stipulates that the registrant can use the list only to remove, or "scrub," a number from their lists. FCC rules specifically prohibit any person (not just a telemarketer) from using any do-not-call list for any purpose other than deleting such numbers from their call lists or sample. Violations are considered an abusive act by the Telemarketing Sales Rule and subject to penalties.

Research industry organizations such as AAPOR (American Association for Public Opinion Research), CASRO (Council of American Survey Research Organizations), CMOR (Council for Marketing and Opinion Research), and the MRA (Marketing Research Association) continue their lobbying efforts in Washington to ensure that telephone survey research remains outside the scope of do-not-call legislation and distinct from telemarketers. In the interest of preserving this distinction, survey organizations are encouraged to define their own written do-not-call policies. By the end of 2006, the Do-Not-Call Registry contained over 132 million landline and wireless residential telephone numbers, according to the FTC's 2007 annual report to Congress. This is over half of the U.S. adult population at the time. Even allowing for multiple numbers per person, household, or both, excluding DNC numbers would necessarily introduce significant coverage bias to telephone surveys and flagging them for special treatment is potentially a violation of FCC and FTC rules.

Linda Piekarski

See also Caller ID; Cell Phone Only Household; Cell Phone Sampling; Do-Not-Call Registries; Number Portability; Predictive Dialing; Telemarketing; Telephone Surveys

Further Readings

Direct Marketing Association. (n.d.). *Telephone Consumer Protection Act*. Retrieved March 5, 2008, from http://www.the-dma.org/guidelines/tcpa.shtml

Federal Communications Commission. (2007). *Unwanted telephone marketing calls*. Retrieved March 5, 2008, from http://www.fcc.gov/cgb/consumerfacts/tcpa.html

Lavrakas, P. J. (2004, May). *Will a perfect storm of cellular forces sink RDD sampling?* Paper presented at the 59th annual American Association for Public Opinion Conference, Phoenix, AZ.

TELEPHONE HOUSEHOLDS

A telephone household is one that has some type of telephone service on which members of the household, in theory, can be reached by an external party, assuming they are called at a time they will answer their telephone. By definition, a telephone survey of the public can include only telephone households in its sampling frame.

In the United States, in 2007, approximately 97% of all households had landline (wired) telephone service, cell (wireless) telephone service, or both. The approximately 3% without service at a given point in time—the nontelephone households—are households that may have service a few months of the year but cannot afford it consistently. These households are disproportionately low-income renters, who live in very rural areas or inner-city poverty areas. Telephone surveys cannot reach (cover) households without any telephone service, and if the topic of the survey is correlated with whether or not a household has telephone service, the telephone survey may suffer from nonnegligible coverage error.

It was not until the 1970s in the United States that telephone service existed in at least 90% of households, although at that time in certain regions of the country less than 80% of households had telephone service. Nowadays the vast majority of households in the United States have both landline and cell telephone service, although reliable and up-to-date statistics on the exact proportions of landline only, cell phone only, and those that have both types of telephone service are not routinely available, especially not at the nonnational level. However, as of late 2007, a federal government study determined that 20% of U.S. households had only cell phone service and approximately 77% had landline service (with most of these also having a cell phone). Survey researchers who choose to sample the public via telephone must pay close attention to the prevalence of telephone households in the geographic areas that are to be sampled. Many challenges exist for telephone survey researchers as of 2008 due in part to (a) the rapid growth of the cell phone only phenomenon, especially among certain demographic segments of the population (e.g., renters and adults under the age of 30), (b) number portability, (c) difficulties in knowing how to properly sample from both the landline and cell frame, especially at the state and local level, and (d) difficulties in knowing how to weight the resulting data that come from both frames.

Paul J. Lavrakas

See also Cell Phone Only Household; Coverage Error; Nontelephone Household; Number Portability; Telephone Penetration; Telephone Surveys

Further Readings

Blumberg, S. J., & Luke, J. V. (2007). *Wireless substitution: Early release of estimates based on data from the National Health Interview Survey, July–December 2006.*

Hyattsville, MD: National Center for Health Statistics. Retrieved April 21, 2008, from http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200705.pdf
Frankel, M. R., Srinath, K. P., Hoaglin, D. C., Battaglia, M. P., Smith, P. J., Wright, R. A., et al. (2003). Adjustments for non-telephone bias in random-digit-dialing surveys. *Statistics in Medicine*, *22*, 1611–1626.
Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E., Japec, L., Lavrakas, P. J., et al. (2008). *Advances in telephone survey methodology*. New York: Wiley.

Telephone Penetration

The term *telephone penetration* refers to the number of households in a given survey area with one or more telephones. Traditionally, this has meant one or more landline or wired telephones, not including cell phones. A major challenge for drawing proper survey samples is ensuring that the sample represents a very high proportion of the population of interest. Traditionally, bias as a result of undercoverage in telephone surveys was credited to households without phones. As time has gone on, the rate of households without phones has been declining, leading to a decline of said bias. In its infancy, telephone interviewing was used only as a method of support for other interviewing techniques, such as face-to-face interviewing. However, by the 1970s the penetration of telephones in U.S. households had exceeded 90%, and this higher telephone penetration resulted in the evolution of telephone interviewing as it has become a centralized and exact data collection method, evolved further through the use of networked computers and computer-assisted telephone interviewing (CATI).

Telephone penetration in the United States has been on the rise ever since the invention of the telephone in 1861. The percentage of households in the United States with telephones increased from less than 40% in 1940, to above 95% as of 2008. Because of coverage issues, telephone surveys, while often regarded as cheaper than face-to-face interviews, lacked respectability for much of the early 20th century. In the first half of the 20th century, telephone ownership was a privilege for those who could afford it. Its use was not common, and ownership was limited to a select group of citizens. For this very reason, telephone directories and telephone surveys were limited to surveys of special populations.

Perhaps one of the most famous examples of the effect of undercoverage on the results of a survey came in 1936, when the *Literary Digest* incorrectly predicted a victory for presidential candidate Alf Landon. Although the survey was conducted by mail, its sample was built from telephone directory listings. In the end, Franklin D. Roosevelt handily defeated Landon, and the general consensus was that poor coverage by the telephone directories used to draw the survey's sample was the cause.

As the number of households equipped with telephones increased, other changes to the survey landscape were also occurring. The increasing costs of face-to-face interviewing and a growing resistance to face-to-face data collection led to the need for researchers to find an alternative that was both affordable and adequate in terms of coverage. Increasing telephone penetration, over 90% by 1970, made telephone surveys a more practical alternative to face-to-face interviewing for obtaining information from a representative sample. As telephone interviewing has become more centralized, interviewing has evolved further through the use of networked computers and CATI.

Although telephone penetration has increased enough over time to make telephone interviewing a viable option for survey research, issues surrounding the potential undercoverage of subgroups in the United States still exist. Research has shown that undercoverage exists in certain demographic subgroups (e.g., Native Americans in the Southwest). Because of this, there is still great concern that the exclusion of non-telephone households from research efforts can lead to the underrepresentation of specific subgroups.

In general, telephone penetration, as it applies to survey research, has focused largely on households with traditional landline telephones. However, increased cell phone usage over time, especially by households who are discontinuing landline service and becoming cell-only, has again threatened the validity of telephone surveys that exclude cell phones from their sampling frames. The increase of cell-only and no-phone households has created a need for researchers to look more closely at the potential biases introduced by excluding cell phone numbers from telephone frames and to develop research strategies that include these numbers.

Excluding cell-only households from random-digit dialing (RDD) sampling frames has become more problematic as the percentage of cell-only households in the United States has risen. As researchers begin to track closely the number of households in the United

States that have become cell-only, the increase in these types of households becomes clearer. By the second half of 2005, the estimated percentage of households in the United States with only a cell phone had reached almost 8%. By early 2008, approximately 20% of Americans could be reached only by cell phone, with an even greater percentage of renters and young adults being cell phone only. Thus, telephone surveys that rely exclusively on landline numbers will have significant undercoverage of these and other portions of the population.

As this trend continues, researchers must attend to the possibility that this growing group, usually eliminated from traditional RDD samples, may display a homogeneity, demographically speaking, that could bias the results of landline RDD surveys. Recent data suggest that those who are cell-only are more likely to be young, nonwhite, and unmarried, with age and marital status subject to the largest undercoverage biases. Further, recent research points to differences between cell-only households and landline households in terms of health and health behavior. Some have dismissed these differences as small and of no threat to the validity of national surveys, but as more people and households convert from landline to cell-only, these biases are likely to increase. Researchers planning to use the traditional landline RDD frame for their telephone surveys will need to attend very closely to the changing penetration of landlines in the population.

David James Roe

See also Bias; Computer-Assisted Telephone Interviewing (CATI); Coverage Error; Cell Phone Only Household; Cell Phone Sampling; Random-Digit Dialing (RDD); Undercoverage

Further Readings

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E. Japec, L., Lavrakas, P. J., et al. (2008). *Advances in telephone survey methodology*. New York: Wiley.

Telephone Surveys

Surveys for which data collection is conducted via a telephone interview represent a major source of all current survey data. Even as Internet surveys have gained greatly in popularity in the past several years, telephone surveys remain a major source of the data gathered for media, marketing, academic, and other types of research. Since the 1970s, the prevalence of telephone interviewing has steadily increased and has surpassed face-to-face interviewing, which had previously been the most commonly used method of conducting survey research. Currently the use of other data collection modes of survey research, particularly Internet surveys, has been increasing, but telephone interviewing still remains a widely used method. This entry discusses systems and techniques used to conduct telephone surveys, the evolution and length of telephone surveys, and the advantages and challenges of such surveys.

Systems and Techniques

Most professional telephone surveys are conducted using a computer-assisted telephone interviewing (CATI) system, which allows, among other things, interviewers to enter respondent information directly into the computer. Questions in CATI can be designed to appear for telephone interviewers one at a time or in blocks of similar questions. This helps to promote a similar interviewing experience for all respondents and limits possible confusion among telephone interviewers. CATI systems also allow *skip patterns*, in which one or more questions are skipped depending on the answers given to previous questions, and other programming such as randomly rotating the order of response options in order to help counterbalance the effect on respondent answers based on the order in which they hear the answer choices.

To further reduce survey error, CATI systems allow for intensive monitoring of interviewers. Interviewers can be monitored for productivity, data quality, and data falsification, among other things. Supervisors are able to monitor interviewers in real time from remote locations. CATI can also improve interviewer productivity by allowing for automatic scheduling of callbacks. This allows interviewers to set up specific times to call back a respondent (often based on a specific respondent request), which helps to increase response rates and productivity.

One of the major benefits of using a CATI system is the ability to easily implement random-digit dialing. This technique is used often in telephone surveys to access a random selection of households within a target population. Random-digit dialing works by allowing area codes and telephone prefixes to be selected in a representative fashion and the remaining digits to be randomly generated by computer. By purposely selecting specific area codes, it is possible to geographically target the geopolitical areas of interest (although the accuracy with which this can be accomplished has decreased considerably due to the effects of number portability and cell phone only households that no longer live in the geopolitical area in which their cell phone was purchased). By specifying telephone prefixes in addition to area codes, many ineligible numbers can be eliminated from the sample even before they are called. Such banks of telephone numbers often include businesses, which many times have different prefixes than do residential numbers. Eliminating these numbers from a sample allows for an increase in overall productivity, as interviewers do not need to spend time calling them.

Evolution

Advances in computer and software technology have allowed the practice of telephone interviewing to evolve over the past few decades. Currently the majority of survey organizations rely on a paperless system that assigns a specific identification number to each telephone number that is part of the sample, that is, to each potential respondent. This allows a record to be kept of every time that number is called and what the outcome of each call was (such as no one was home, a busy signal was received, an interview was completed, etc.). It also enables a record to be kept of the date and time that each call to a number was placed, which helps to ensure that numbers in the unresolved sample are called back at the appropriate time (such as calling back a respondent who is available only in the afternoons).

Length

Telephone surveys vary greatly in length. Many news and election-related polls can be on the shorter side: approximately 2 to 5 minutes. Surveys employed for academic research can range up to 45 minutes or more, and market research surveys are often somewhere in between. Longer surveys tend to elicit a lower participation rate due to respondent fatigue. However, many media polls and marketing surveys, for example, are not as concerned with the response rate of surveys and instead focus on obtaining a specific number of completed interviews without regard for the number of telephone numbers at which no person was reached or for which the person at the number refused to participate. Other surveys, in particular government and academic surveys, are more concerned with these aspects because they need to obtain results that are more certain to be representative of the population of interest.

Advantages

Telephone surveys have many advantages over face-to-face interviewing, mail surveys, and Internet surveys. When compared to face-to-face and mail surveys, telephone surveys can be completed much more quickly and cheaply, which is almost always a major consideration in the field of survey research. Although telephone surveys are still more expensive than Internet surveys, they have the advantage that they are much better able to represent the general population, despite the challenges posed by cell phone only households in the United States. Even though Internet access and use have been on the rise over the past several years, the percentage of U.S. households that have such access is still not as high as the percentage that can be reached by telephone (97% in 2008). Therefore, in surveys that attempt to contact members of a large or diverse population, the representativeness of the data is questionable when Internet surveys are employed instead of telephone surveys. Furthermore, surveys that use interviewers, such as telephone surveys, will almost always elicit higher response rates, all other factors being equal, than surveys that do not use interviewers, such as Internet surveys.

Challenges

The telephone survey represents a major way that a large population can be accurately surveyed, but it still possesses many challenges to survey accuracy. Response rates have been declining as the use of caller ID and cellular telephones has been on the rise, which represents a major problem for the industry. As response rates decline, the accuracy of the data becomes more questionable. For example, if only a small portion of the selected sample participates in a survey, it could mean that those people who were sampled but did not participate are, in some way, inherently different from those who do participate and would therefore give different responses to the interview questions. The end result would be that the data would be skewed toward a certain set of opinions due to nonresponse bias. Coverage rates in landline telephone surveys also have decreased to below 80% as of 2008.

Innovations such as caller ID and privacy manager allow potential respondents to screen out calls before they even find out the purpose of the call. Telemarketers are generally thought to be a compounding factor because so many households receive a large number of telephone solicitations from telemarketers, and potential respondents begin to assume that any telephone number appearing on caller ID that is not recognizable is likely to be a telemarketer. Another contributor to the decline in telephone survey response rates are the popular do-not-call lists. Created to reduce the number of unwanted telemarketing calls, many people assume that if they are on such a list, they should not be called by anyone whom they do not personally know. Even though such lists do not apply to legitimate survey research organizations, many potential respondents do not make such distinctions. Additionally, the increased use of the telephone to conduct surveys over the past few decades has contributed to the decline in response rates. As people become used to receiving solicitations for their opinions, it becomes easier to refuse those requests.

Finally, the use of SUGing and FRUGing has also contributed to the decline in response rates for telephone surveys. SUGing (Soliciting Under the Guise of survey research) and FRUGing (Fund-Raising Under the Guise of survey research) is done by organizations that attempt to give the purpose of the calls greater legitimacy by claiming to only be interested in surveying opinions. This tactic quickly degenerates into a sales or fund-raising call but leaves the lasting impression on the respondent that telephone surveys often are merely a disguised way of selling products or raising money.

Mary Outwater

See also Caller ID; Cell Phone Only Household; Cell Phone Sampling; Computer-Assisted Telephone Interviewing (CATI); Do-Not-Call Registries; FRUGing; Interviewer Monitoring; List-Assisted Sampling; Nonresponse Bias; Number Portability; Random-Digit Dialing (RDD); Response Rates; SUGing

Further Readings

Groves, R. M., Biemer, P. P., Lyberg, L.E., Massey, J. T., Nicholls, W. L., II, Waksberg, J. (Eds.). (1988). *Telephone survey methodology*. New York: Wiley.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Eds.). (2004). *Survey methodology*. New York: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision*. Newbury Park, CA: Sage.

Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E., Japec, L., Lavrakas, P. J., et al. (2008). *Advances in telephone survey methodology*. New York: Wiley.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling and data analysis* (3rd ed.). Thousand Oaks, CA: Sage.

TELESCOPING

Telescoping describes a phenomenon that threatens the validity of self-reported dates, durations, and frequencies of events. Respondents often are asked in surveys to retrospectively report when something occurred, how long something lasted, or how often something happened within a certain time period. For example, health surveys often ask respondents the date of the last time, how often, or how many days they were hospitalized during the last calendar year. Answering this type of question requires the respondent to remember exact dates and temporal sequences and to determine whether an event happened within a certain time period. At this stage of the response process, dates or events can be forgotten entirely or "telescoped" forward or backward. While forgetting describes not remembering an event at all, telescoping focuses on errors made by incorrectly dating events that were recalled.

Survey researchers distinguish between two types of telescoping: *forward* and *backward*. Forward telescoping occurs when an event is erroneously remembered as having occurred more recently than it actually did. A backward telescoped event is erroneously remembered as having occurred earlier than its actual date. In general, empirical data show that forward telescoping is more likely to occur than backward telescoping.

Why telescoping occurs is not fully understood. Two main theories have emerged in the literature: the time compression and variance theories. However, these theories explain only parts of the phenomenon. The *time compression theory* focuses only on explaining forward telescoping, arguing that telescoping occurs because of a subjective distortion of the time line. Time is compressed when respondents perceive that events happened more recently than they actually did or a time period seems shorter than the true length of time. This theory also hypothesizes that forward telescoping decreases as the length of the reference period increases. Empirical findings testing this hypothesis, however, have been mixed. Variance theory uses the uncertainty in one's memory about the time of an event as an explanation for telescoping. The theory argues that uncertainty about the timing of an event increases as the elapsed time from the event to when the question is asked expands, explaining both forward and backward telescoping.

Several methods are used to reduce the amount of telescoping. First, "landmark events" can be used to clearly mark the beginning of the reference period. Landmark events are defined as personal or public events that are meaningful and highly salient to the respondent and can therefore provide a temporal structure of events; for example, something happened before or after a car accident. Personally meaningful events are better encoded in autobiographical memory than public events and therefore appear to be better landmark events. These events limit the reference period, provide a temporal structure to the events, and increase accuracy of reports. Bounded interviews are also used to reduce the incidence of telescoping. Used most frequently in panel surveys, bounded interviews permit the interviewer to remind the respondent of his or her reports in the previous interview or to check for overlaps between the current report of events and previously reported events. This technique improves report accuracy by eliminating forward telescoping of previously reported events. Finally, decomposition of a question into several more specific questions has been used. Research has shown that decomposition improves reporting only if the behavior is irregular and dissimilar. Otherwise, decomposing the question can lead to less accurate reports. Thus, the effectiveness of this technique varies over the population, and the use of this strategy should be carefully assessed.

Sonja Ziniel

See also Behavioral Question; Bounding; Measurement Error; Overreporting

Further Readings

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Temporary Dispositions

Temporary disposition codes are used to record the outcomes of specific contact attempts during a survey that are not final dispositions, and provide survey researchers with the status of each unit or case within the sampling pool at any given point in the field period. Temporary disposition codes provide a record of what happened during each contact attempt prior to the case ending in some final disposition and, as such, provide survey researchers with the "history" of each active case in a survey sample. Temporary dispositions function as an important quality assurance component in a survey—regardless of the mode in which the survey is conducted. They also serve as "paradata" in some methodological research studies. However, the primary purpose of temporary dispositions is to assist researchers in controlling the sampling pool during the field period.

Temporary dispositions usually are tracked through the use of an extensive system of numeric codes or categories that are assigned to each element in the sampling pool once the field period of the survey has begun. Common temporary sample dispositions include the following:

- No one at home/answering
- Busy signal (telephone survey only)
- Fast busy (telephone survey only)
- Callback
- Privacy manager (telephone survey only)
- Unable to participate
- Unavailable respondent
- Household refusal (a temporary disposition if refusal conversions are planned)
- Respondent refusal (a temporary disposition if refusal conversions are planned)

Temporary disposition codes may also be matched with "action codes" that take into consideration the *status* of a case at any given point in the field period and lead logically to what the next action on the case should be. Examples of these types of action codes include *maximum number of call attempts* and *supervisor review*. Although these action codes are important in managing a survey sample, they should not be used as temporary disposition codes.

Temporary dispositions may change often during the field period of a survey—usually as often as the status of each case in the sample changes or as interviewers work cases in the sample. For example, the temporary disposition code of each telephone number in the sample for a telephone survey is updated after every call that is made to the number by an interviewer. In the case of a mail survey, sample dispositions may be updated as completed survey questionnaires are returned to researchers by respondents or as the postal service brings questionnaires "returned to sender" or "post office return" in the case of incorrect addresses or respondents who have moved. In an Internet survey, sample dispositions may be updated as email invitations are sent to individuals in the sampling pool, as email messages are returned to the sender after not being able to be delivered (in the case of an incorrect email address), as respondents log in to complete the Web survey, and as respondents complete the questionnaires.

Currently there is no standardized set of temporary disposition codes, and many survey firms develop their own systems. Although this is not a problem, it is important that the system of temporary codes used by an organization be compatible with the standard definitions of final case disposition codes that have been developed by survey-related professional organizations such as AAPOR (American Association for Public Opinion Research).

Matthew Courser

See also Dispositions; Final Dispositions; Paradata; Standard Definitions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Test–Retest Reliability

Test–retest reliability is a statistical technique used to estimate components of measurement error by repeating the measurement process on the same subjects, under conditions as similar as possible, and comparing the observations. The term *reliability* in this context refers to the precision of the measurement (i.e., small variability in the observations that would be made on the same subject on different occasions) but is not concerned with the potential existence of bias.

In the context of surveys, test–retest is usually in the form of an interview–reinterview procedure, where the survey instrument is administered on multiple occasions (usually twice), and the responses on these occasions are compared.

Ideally the reinterview (henceforth referred to as T_2) should exactly reproduce the conditions at the original interview (T_1). Unfortunately, learning, recall bias, and true changes may have occurred since the original interview, all leading to T_2 not matching T_1 answers.

Which measurement error components are of interest may affect certain decisions, such as whether to use raw data (if the focus is on response error alone) or edited data and imputed data (if editing and imputation errors are of interest). Oftentimes raw responses need to undergo light editing for practical reasons. Reliability measures (discussed later in this entry) may be calculated for individual questions to identify ones that may be poorly worded or others that may need to be followed by probing. Also, reliability measures may be calculated for population subgroups or by type of interviewer (e.g., experienced vs. novice) to detect problems and improve precision.

Further considerations in the design of a test–retest study include considering whether the interview–reinterview sample should be embedded in the main survey sample or exist as an additional, separate sample. An embedded sample is obviously cheaper, but sometimes a separate sample may be more advantageous. Another issue to consider is using the same interviewer at T_2 or not. Using the same interviewer in part of the sample and assigning different interviewers to the rest may help in assessing interviewer effects. The duration of time between T_1 and T_2 (the test–retest "lag") often has an effect on the consistency between the interviews. Thus, limiting to a certain time range may be an important consideration.

Furthermore, it is important to properly inform the respondents of the need for the reinterview, because respondents' perceptions of its necessity may affect the quality of the data they provide. After the reinterview is completed, additional special questions may be asked of respondents, such as how much they remembered from the original interview and what effect it may have had on the responses. The interviewers themselves may be debriefed on what they observed in the reinterview.

Considerations in the Analysis of Test–Retest Studies

The analyses may use information gathered at T_1 on T_2 nonrespondents for the purpose of making nonresponse adjustments, particularly if refusal to T_2 was related to questions asked at T_1. Reliability may depend on the interviewer and on interactions between interviewers and respondents (e.g., dependence on interviewer and respondent being of same gender or background may affect reliability). Further, if some of the respondents were interviewed by the same interviewer at T_1 as they were at T_2, whereas others were not, relationships between this factor and reliability can be analyzed. Because the time lag between T_1 and T_2 may have an effect on recall and learning effects, analysis of the dependence of the agreement between the two interviews on the lag may be warranted. The relative duration of the T_2 interview to that of T_1 may be related to learning (e.g., shortening of duration) and also should be explored.

Measures of Test–Retest Reliability

There are a number of reliability measures proposed in the literature for nominal, ordinal, and continuous (interval and ratio) responses. Although some of these measures are widely used, there is no general consensus which ones are best.

A raw measure of reliability of any single item is the proportion (%) of agreement. Consider the case of a binary variable having levels Y and N as shown in Table 1.

The proportion of agreement is $p_0 = p_{yy} + p_{nn}$. Because part of this agreement could be by chance alone, a chance-corrected measure called kappa (κ), as proposed by J. Cohen in 1960, is defined by $\kappa = (p_0 - p_e)/(1 - p_e)$ where p_e is the probability of chance agreement: $p_e = p_{y+}p_{+y} + p_{n+}p_{+n}$. Extension to the general nominal variable case is straightforward. For categorical responses, Cohen's kappa is the most commonly used measure in reliability studies. Nevertheless, there has been criticism of some of kappa's undesirable properties (e.g., when the marginal distributions of the prevalence are asymmetric). Weighted kappa is a version of Cohen's kappa for ordinal responses, where differences between the levels of the response variable are weighted, based on judgment of the magnitude of the discrepancy. For continuous responses, a measure called the concordance correlation coefficient was proposed by L. Lin in 1989.

Table 1 Test–retest proportion of agreement with a binary variable

		T_1 Y	T_1 N	Row total
T_2	Y	p_{yy}	p_{yn}	p_{y+}
	N	p_{ny}	p_{nn}	p_{n+}
Column total		p_{+y}	p_{+n}	1

L. Pritzker and R. Hansen proposed another measure called the index of inconsistency (IOI). Its complement, $1 - \text{IOI}$, is referred to as the index of reliability. The IOI is defined as the ratio of the measurement variance to the total variance, and thus can be viewed as an intracluster correlation. (Of note: Some measures defined for continuous responses can also be used for dichotomous or ordinal data.) Finally, various types of models have also been used to make inference on reliability. While modeling involves making certain distributional assumptions, it allows for adjusting for covariates.

Moshe Feder

See also Measurement Error; Probing; Raw Data; Reinterview; Reliability

Further Readings

Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics*, *52*, 695–702.

Blair, J., & Sudman, S. (1996). The survey reinterview: Respondent perceptions and response strategies. *Journal of Official Statistics*, *12*, 421–426.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Falorsi, P. D., Pallara, A., & Russo, A. (2003, November 17–19). *Measuring response errors in censuses and surveys through reinterview data having a different measurement error variance*. Paper presented at the

Federal Committee on Statistical Methodology Research conference, Arlington, VA. Retrieved April 21, 2008, from http://www.fcsm.gov/03papers/falorsi_pallara_russo.pdf

Fleiss, J. L. (Ed.). (1981). The measurement of interrater agreement. In *Statistical methods for rates and proportions* (pp. 212–236). New York: Wiley.

Forsman, G., & Schreiner, I. (1991). The design and analysis of reinterview: An overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 279–301). New York: Wiley.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility *Biometrics*, *45*, 255–268.

Pritzker, L., & Hanson, R. (1962). Measurement errors in the 1960 census of population. In *Proceedings of the Section on Social Statistics* (pp. 80–89). Washington, DC: American Statistical Association.

Yawn, B. P., & Wollan, P. (2005). Interrater reliability: Completing the methods description in medical records review studies. *American Journal of Epidemiology*, *161*, 974–977.

Third-Person Effect

The *third-person effect* is a term that refers to the documented belief held by many people that mass communication has different and greater effects on others than on themselves, and because of this perception, some of these people will support certain policies and actions based upon the presumed effect on others. The phenomenon has been linked to public opinion research, and it often is studied through survey research methods.

Background and Theoretical Origins

What began as an eclectic litany of recollections and ruminations accumulated over nearly a lifetime of one scholar's experience—supplemented by little formal data or analysis—blossomed into a fertile site of rigorous interdisciplinary scholarship.

In his seminal work in the early 1980s, W. Phillips Davison relayed several anecdotes of how different people in different circumstances estimated different presumed impacts of the same messages. Davison later reported that he "didn't really want to write the article," in part because he thought the phenomenon was of "minor theoretical significance" and his observations were based on sketchy data. Nevertheless, his observations were considered intriguing by many others who read and elaborated on his work, and in 2004 the third-person effect was named one of the most popular communication-theory frameworks of the early 21st century.

Davison explained the third-person effect in the following terms:

- People will tend to overestimate the influence that mass communications have on the attitudes and behaviors of others.
- People will expect the communication to have a greater effect on others than on themselves.
- Whether or not these individuals are among the *ostensible* audience for the message, the impact that they expect this communication to have on others may lead them to take some action.

Davison went on to explain that there are two ways in which the notion of a "third person" can be interpreted. First, individuals often believe that people like "me" or "you" will not be impacted by communications as much as "them," that is, the third persons. Second, some individuals who themselves are not members of the ostensible (or intended) audience nonetheless are concerned about the presumed effects of messages on the ostensible audience. These third persons, especially if in positions of authority, are driven by this presumption of effects to make what could be characterized as paternalistic decisions about the fate of the messages and the rights of members of the ostensible audience to be exposed to the messages.

Research on the Third-Person Effect

Since Davison's original articulation of the third-person effect, considerable scholarly effort has been invested in attempts to provide a suitable theoretical context and explanation for Davison's observations and insights. Other efforts have focused on methodological issues relating to study and instrument design. For example, Richard Perloff has written several broad and cogent analyses of theoretical and methodological conundrums associated with formal empirical tests of Davison's propositions.

Over the past 25 years, the third-person effect has been explained by, and linked to, a variety of established social psychological theories and models, including attribution, social comparison, social desirability, social distance, unrealistic optimism, symbolic interactionism, and self-categorization, among others. The third-person effect has also been connected to the larger scholarly body of public opinion theory and research, including such phenomena as spiral of silence and pluralistic ignorance. Additional theoretical attention has been paid to the role that message variables may play in contributing to the third-person effect. Evidence is mixed across studies, but there is some indication that the desirability of a message, that is, whether it is pro- or anti-social, can influence the nature and magnitude of the third-person effect. In addition, some studies have found that certain messages can elicit a "reverse" third-person effect, that is, a situation in which individuals report a greater effect of the message on one's self than on others. Another key consideration in third-person research has to do with the nature of the "other" and his or her relation to, and psychological distance from, the self.

Turning to issues more methodological in nature, early formal tests of third-person propositions tended to employ lab-based experimental designs, but the phenomenon has been studied extensively with survey research methodology as well. Results across studies indicate that the third-person effect occurs independently of the methodological approach used to study it. Several studies have investigated whether the phenomenon is simply a methodological artifact relating to question wording, question order, or study design (e.g., whether a between- or within-subjects design is employed). Results of these studies indicate that the third-person effect is more than a mere methodological quirk.

What originally appeared to be an assortment of observations of seemingly minor theoretical significance subsequently has achieved the status of a vibrant media-effects approach that has important implications for survey research and public opinion theory.

Charles T. Salmon

See also Experimental Design; Public Opinion; Public Opinion Research; Question Order Effects; Social Desirability; Spiral of Silence

Further Readings

Andsager, J. L., & White, H. A. (2007). *Self versus others: Media, messages, and the third-person effect*. Mahwah, NJ: Lawrence Erlbaum.

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, *47*(1), 1–15.

Gunther, A. C., Perloff, R. M., & Tsfati, Y. (2008). Public opinion and the third-person effect. In W. Donsbach & M. Traugott (Eds.), *Handbook of public opinion*. Thousand Oaks, CA: Sage.

Perloff, R. M. (in press). Mass media, social perception and the third-person effect. In J. Bryant & M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (3rd ed.). New York: Taylor & Francis.

Sun, Y., Pan, Z., & Shen, L. (in press). Understanding the third-person perception: Evidence from a meta-analysis. *Journal of Communication*.

Topic Saliency

The saliency of a topic—that is, its importance or relevance to potential respondents—can affect response patterns to surveys. There are several explanations as to how and why topic saliency affects response rates. First, if a potential respondent believes the topic to be important, he or she may be more likely to rationalize incurring the costs of responding to the survey. Second, responding to a survey topic that is of personal interest may have intrinsic rewards, such as providing an opportunity to exhibit one's knowledge or share one's opinion. Third, responding to a survey about a salient topic may be motivated by perceived direct benefits. Survey participation may be viewed as an opportunity to advance one's own needs, interests, or agenda. All of these explanations may apply to explain why a single respondent or respondents are more apt to complete a survey about a salient topic.

Research suggests that people are more likely to respond to a survey if the topic is of interest to them. For example, teachers are more likely than nonteachers to respond to, and cooperate with, a survey about education and schools; senior citizens are more likely than younger adults to respond to a survey about Medicare and health.

In addition to its impact on survey participation, topic saliency is important for attitude formation and response retrieval. Theories about the representation of attitudes in memory suggest that attitudes reported by the respondent as being more important or as being more strongly held are also more stable over time. Attitudes about salient topics require less cognitive effort to recall, resulting in attitude responses that are more stable over time, more stable when presented with counterarguments, and more consistent with other attitudes and considerations.

There is great debate about whether attitudes are saved in memory and retrieved when the situation or the survey question arises (the online model of attitude formation), or whether attitudes are continually constructed from multiple considerations that are sampled each time they are needed (the memory model). The online model implies a continuum ranging from nonattitudes to "true" attitudes, where attitude stability and consistency are partly determined by topic saliency. The memory model implies that attitude formation is a stochastic process subject to some variability where the respondent samples considerations off the "top-of-the-head" when asked a survey question. In this model, more salient topics result in a response distribution for each individual that is more tightly clustered and therefore also more stable.

Whether the attitude is "true" or constructed on the spot, attitudes about salient topics are often considered to be more resistant to differences in questionnaire form. However, the evidence is mixed and depends highly on how topic saliency is measured. Self-reports of attitude importance, certainty, or strength are more resistant to certain questionnaire design features. When salience is measured by interest in politics or by the cognitive accessibility of those attitudes, evidence is tenuous on whether topic salience is related to questionnaire form or other survey response patterns.

Topic saliency has implications for survey operations, countering survey nonresponse, questionnaire design, and analysis. People are more likely to cooperate with, and respond more quickly to, a survey if the survey topic is of interest to them, suggesting a need to compensate for the potential bias this effect may cause by using a multiple contact strategy. Questionnaire designers and analysts need to consider the implications of their question form for their response distribution and thus for the interpretation of the results.

Shelley Boulianne and Danna Basson

See also Attitudes; Leverage-Saliency Theory; Nonattitude; Nonresponse; Questionnaire Design; Response Rates; Saliency

Further Readings

Bassili, J. N., & Krosnick, J. A. (2000). Do strength-related attitude properties determine susceptibility to response effects? New evidence using response latency, attitude extremity, and aggregate indices. *Political Psychology*, *21*(1), 107–134.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *68*(1), 2–31.

Total Design Method (TDM)

The total design method (TDM) is an approach to obtaining response to surveys. Under this approach, social exchange theory is used to identify ways to improve the quantity and quality of survey response by organizing the data collection process in a way that increases trust that the *rewards* of responding will be seen by the respondents as outweighing the *costs* of doing so. The TDM was developed by Don A. Dillman, in 1978, as a general framework for designing both mail and telephone surveys, but it is most identified with developing and implementing surveys by mail. In recent years, it has been recast as the tailored design method and applied to the design of Internet and mixed-mode surveys as well as postal surveys.

Elements of the Total Design Method

The original TDM consisted of two parts. The first was to identify each aspect of the survey process likely to affect either the quality or quantity of responses and to shape them in ways that would improve response. The second part was aimed at organizing the survey efforts so that the design intentions were carried out in complete detail.

The problem that the TDM was designed to solve was that much of the research literature on mail survey design emphasized the individual influence of single techniques—from sending multiple contacts to placing real stamps on return envelopes—without focusing on combined overall effects aimed at achieving the best possible response from respondents. Combining techniques into an overall approach to data collection, focusing on both elements and their temporal interconnections, raised issues of compatibility and of how the use of some techniques might need to be reshaped to be compatible with other response-inducing techniques.

To make decisions on how to combine multiple techniques, social exchange theory was utilized as a conceptual framework. The behavioral assumption implied by this framework is that people's actions are typically motivated by their expectations that in the long run, the rewards for taking action will outweigh the costs of doing so. The process of sending a questionnaire to sampled individuals, persuading them to complete it in an accurate manner, and return it were viewed as a special case of social exchange.

Social exchange differs in significant ways from economic exchange. Social exchange involves diffuse obligations, whereby one person does something in anticipation of the likelihood that the other person will do something in response that will benefit the respondent or others. In addition, the reciprocal obligation is not something that can be bargained with. With respect to social exchange in survey research, it is left up to the potential respondent to take action based upon what the sender of the questionnaire has already done. Social exchange contrasts in significant ways with economic exchange, for which people involved in a transaction typically agree on a price before the transaction occurs and may bargain on that price prior to deciding whether to participate in it. Social exchange does not generally involve explicit bargaining.

Obtaining a Positive Response From Respondents

In the design of mail surveys, three different factors are subject to design actions aimed toward obtaining a positive response from respondents: (1) rewards, (2) costs to the respondent, and (3) the trust of the respondent that, in the long run, the rewards for completing the survey will outweigh its costs.

Rewards

Normally, rewards for completing a questionnaire are small. Social factors that have reward value may include explaining why completing a questionnaire is helpful to solving a problem that the respondent considers important. Writing something supportive of the values held by the survey respondent may also have positive reward value. In addition, developing and ordering questions in ways that make them interesting to the respondent and easier to understand may have reward value, as does expressing appreciation for the respondent's help.

Costs

The second factor important in social exchange is the costs to the respondent. Social costs can be lowered. One way to do this is by making sure that letters are worded so that they do not subordinate the respondent, psychologically, to the study sponsor (e.g., "I'm doing this study to help people like yourself improve your eating habits"). Embarrassing the respondents by creating questions that are difficult, require knowledge the respondent does not have, or create anxiety for other reasons may incur social costs to respondents. Costs that flow from consideration of how the respondent is treated go far beyond factors normally thought of as having economic cost.

Trust

The third important aspect of the social exchange equation is trust on the part of the respondent that, in the long run, rewards for completing a questionnaire are likely to exceed the costs incurred by that action. Respondents need to feel that promises made by the survey sponsor (e.g., "to make public officials aware of the results of the survey") will be carried out. Sponsorship by an organization that is viewed as legitimate by the person asked to respond is an important feature that may create trust that research findings from a survey will be helpful to someone. This helps to explain why surveys sent by the government typically get higher response rates than those sent by market research organizations. The sending of token cash incentives of a few dollars with a request, which has been found to be quite effective for improving survey response rates, may stem from the creation of trust that a survey is important. Inclusion of token noncontingent incentives ahead of time conveys both the idea that the survey is important and trust that the respondent will consider the request. Research has shown that contingent cash payments, even sizable ones, offered in return for returning a completed questionnaire are far less effective than token cash incentives sent ahead of time. This difference provides strong evidence of the social, as opposed to economic, nature of the questionnaire completion process.

Based on the social exchange framework, a series of specific recommendations for designing mail surveys was developed. Features included (a) ordering questions in a particular manner from being of greater to lesser interest to respondents, (b) printing questionnaires as booklets small enough to fit in regular business mail envelopes when folded, (c) particular sizes of envelopes, (d) letters with specified content, (e) personalization of those letters through individually typed names and addresses, (f) a timed sequence of four contacts (questionnaire, postcard reminder, replacement questionnaire, another replacement questionnaire), and (g) stamped return envelopes. Application of these procedures in concert produced mail survey response rates averaging more than 70% in the 1970s from a wide variety of survey populations.

Uses of the TDM

Although the TDM is also applied to telephone surveys, the window of opportunity for applying social exchange principles, except for the use of advance letters, is narrow. In addition, differences in response rates between telephone surveys that use the TDM approach and those that do not is much less than for mail surveys. Thus, the TDM has been identified primarily as a mail survey methodology.

TDM principles have been applied successfully to the design of thousands of mail surveys from the 1970s through the 1990s. Surveys acknowledging use of the total design framework have been conducted throughout the United States by government, education, and private organizations. Published surveys have also been conducted in many countries throughout the world, such as New Zealand, Japan, China, Australia, the Netherlands, Spain, the United Kingdom, Israel, and Germany.

Critiques

One of the effects of publication of the TDM was to shift the emphasis away from looking for a magic bullet that might improve response rates toward providing a template, which, if followed carefully, was likely to produce high response rates. Its use also has shown that mail response rates of the public can often rival or even surpass those obtained in telephone surveys; this was true especially for surveys conducted in the 1980s and 1990s. Thus, the overall effect was to give legitimacy to mail surveys as a way of undertaking serious research.

It also became apparent that the TDM was not an adequate design for mail surveys in certain respects. First, the TDM, as originally conceived, was focused heavily on improving response rates. Although response rates are important, they remain one step removed from nonresponse error (e.g., differences between respondents and nonrespondents that are relevant to the study objectives). Although detailed attention was given to survey measurement and the writing of unbiased questions in the original TDM, comparable interest was focused on organizing questions in a way that would entice recipients of questionnaires to respond. Coverage error was, and remains, the major barrier to the effective use of mail surveys for general public populations because of the lack of adequate sampling frames for drawing samples. In addition, satisfactory selection procedures for randomly choosing particular respondents within households to represent the general public remain more of a concern for self-administered mail surveys than for either personal or telephone surveys.

Second, the TDM also had a one-size-fits-all character (e.g., the same procedures were advocated for use in all survey situations). The need to modify specific procedures for special situations and populations was mostly ignored. Such challenges range from the completion of time-sensitive diaries (e.g., Nielsen's 7-day TV ratings diary), in which it makes no sense to use the 2-month implementation cycle described as a standard TDM procedure, to business surveys that have to be passed from person to person to be completed and thus require a far longer time frame.

Third, concerns about encouraging response selectivity were raised by the strong focus on ordering questions from most to least interesting and on giving corresponding attention to the use of letters explaining the survey topic and why it was of importance. That is, appeal to respondent interests in the survey topic might discourage those with less interest in the survey content. Thus, concern was raised as to whether use of the TDM would increase nonresponse error rather than decrease it.

In addition, the TDM was bound by the technology of the times. Advocating the insertion of individual names and addresses into the salutation space of preprinted letters represented a significant advance forward in personalization in the 1970s. However, it was a typewriter-based technology that is as quaint today as it was innovative and time-consuming at that time.

The TDM also exhibited a bias toward university-sponsored studies and rural as well as state population surveys associated with the location of the university where the TDM was initially developed. Evidence was lacking that it would work for national surveys, particularly those of residents in large cities. Although such studies were eventually done, response tended to be lower and perhaps less subject to some of the exchange principles advocated in the original TDM.

Replacement of the TDM With Tailored Design

In 2000, some 20 years after its original development and publication, Dillman replaced the total design method with the tailored design method. The revised approach retained the social exchange framework for design, but it began with the recognition that the full application of social exchange would suggest using different procedures to obtain response from different populations, on different topics and in different survey situations, as well as sponsorship, which is consistent with leverage-saliency theory. Thus, an attempt was made to more specifically identify issues that may increase rewards, decrease costs, or encourage trust with specific populations or even portions of a particular population (e.g., small as opposed to large business owners).

In addition, tailored design was connected explicitly to the four cornerstones of total survey error: sampling error, coverage error, measurement error, and nonresponse error. For example, relatively less emphasis was placed on response rates and more on finding ways to ensure results did not exhibit nonresponse error, by methods such as using token cash incentives in advance, which are more effective with younger people (who are less likely to respond to surveys) than they are with older respondents (who have a greater propensity to answer questionnaires). In addition, more emphasis was placed on mixed-mode surveys, which provide a way of getting respondents who cannot or will not respond by one mode to respond via another one.

The tailored design recognizes that methods inappropriate or less effective with some populations may be acceptable or effective with others. For example, while six or seven contacts made within a 2- or 3-week period would be appropriate when trying to obtain television or radio listening for a specific set of days, that number of contacts in such a short period of time would not be appropriate for a lengthy business survey that required involving several people in order to get a complete response. Also, whereas government surveys carry great legitimacy and obtain relatively high response rates without the use of cash incentives, private-sector surveys make the use of such incentives more appropriate. The variety of situations calling for design variations is quite large, ranging from employee surveys sent through internal company mail (for which there is strong employer encouragement to respond) to election surveys that use voter registration lists and require focusing contacts on a deadline date.

The development of the tailored design, or the "new TDM" as it is sometimes called, also reflected a rapidly growing science based on factors that influence responses to survey questions. Research that began in the 1990s made it clear that people who respond to mail surveys are influenced by the graphical layout of questionnaires as well as the use of symbols and numbers, and that their reactions to these features of construction are influenced by behavioral principles such as the gestalt laws of psychology. Whereas the original TDM included a focus on visual layout, the scientific underpinnings of those principles had not yet been demonstrated or methodically tested. Another aspect of the science base that provided a basis for tailored design was nearly two decades of research on question order and context affects. Thus, measurement principles developed for tailored design gave much greater importance to what questions were to be asked and the order in which they needed to be asked as well as how they were to be displayed on pages, while giving somewhat less interest to what questions would appeal most to respondents.

Development of a tailored design perspective was facilitated by the development of information technologies that make it possible to treat different sample units in different ways, efficiently and accurately. Active monitoring of response and the timely tailoring of specific follow-up letters to different types of respondents were not feasible in the 1970s.

Internet surveys have now been encompassed within the tailored design perspective. Because mail and Web surveys are both self-administered as well as visual, they share many commonalities. Whereas the technological interface between respondent and response mechanisms clearly distinguishes Web from mail, its similar exchange processes appear to occur between respondent and the questionnaire and its sponsor. Both mail and the Internet face the challenge of not having an interviewer who can provide additional explanations for what questions mean and ask probing follow-up questions that are responsive to answers and comments just offered by the respondent.

During the past decade, evidence has accumulated that fundamentally different communication processes occur in visual (Web and mail) versus aural (telephone) surveys. Telephone surveys rely mostly on words for communication, which are given a limited amount of additional meaning by interviewer characteristics that permeate the delivery of those words. However, mail and Web surveys depend not only on words to communicate question meaning but also on symbols, numbers, and graphics that may give meaning to verbal language, quite apart from words themselves. Experimental research published since 2000 makes it quite evident that answers to survey questions and the completion process are influenced by how scales are visually displayed. Identification of these differences in respondent answers now raises important questions about which visual formats translate most effectively between aural and visual modes of conducting surveys.

The original TDM was developed at a time when mail was by far the least expensive survey mode, costing far less to implement than the alternatives of telephone or face-to-face interviews. The attention to detail encouraged by the TDM raised mail costs significantly, but they remained low compared to the face-to-face and telephone alternatives. In the new TDM era, mail surveys are now considered a high-cost methodology, fairly equivalent in cost to telephone interviews but far more expensive than Internet surveys once the fixed costs of creating Internet surveys have been met. In addition, declining response rates to the telephone and low response to and the limited coverage of Web surveys have provided strong encouragement for mixing modes. Thus, in the early 21st century a new feature of tailored design receiving much attention is how to use it to obtain as many responses as possible by Web and to supplement those responses with surveys by mail or telephone of respondents who lack access to, or will not respond over, the Internet.

One of the practical outcomes of development of the original TDM and now tailored design is that the attention of surveyors is drawn to the fact that responding to surveys involves communicating effectively with potential respondents through all aspects of the survey design process and not just survey questions. It is a process in which both the surveyors and the respondent take into account the expectations of the other. Its effectiveness as a guide for designing surveys has gained importance, as surveyors have recognized the explicitly voluntary nature of surveys, and seems likely to continue to influence survey design throughout the world.

Don A. Dillman

See also Advance Letter; Contingent Incentives; Coverage Error; Economic Exchange Theory; Gestalt Psychology; Incentives; Internet Surveys; Leverage-Saliency Theory; Mail Questionnaire; Mail Survey; Noncontingent Incentives; Nonresponse Error; Response Rates; Social Exchange Theory; Total Survey Error (TSE); Visual Communication; Within-Unit Coverage

Further Readings

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.
Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, *17*, 225–249.
Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method*. New York: Wiley.
Dillman, D. A. (2007). *Mail and internet surveys* (2nd ed.). New York: Wiley.

Total Survey Error (TSE)

Total survey error (TSE) is a term that is used to refer to all sources of bias (systematic error) and variance (random error) that may affect the validity (accuracy) of survey data. Total error in surveys can be conceptualized and categorized in many ways. One traditional approach is dividing total error into sources of sampling error and sources of nonsampling error. Another categorization is dividing it between coverage error, sampling error, nonresponse error, and measurement error. A more modern approach is to group various sources of error into the classes of *representation* and *measurement*. This entry provides a big picture perspective on all of the major types of error that occur in surveys and thus comprise total survey error.

Unfortunately, there is no such thing as a survey without error. Nevertheless, survey methodologists and survey practitioners aim for the most accurate surveys that can be conducted given the finite budget available to fund them. The quality of a survey statistic such as a mean, a percentage, or a correlation coefficient is assessed by multiple criteria: the timeliness of reporting, the relevance of the findings, the credibility of researchers and results, and the accuracy of the estimates—just to mention a few. Among those criteria the accuracy of the estimate is not necessarily the most important one. However, the accuracy is a dimension of the overall survey quality for which survey methodology offers a wide range of guidelines and instructions. Also, standard measures for the magnitude of the accuracy are available. The accuracy of a survey statistic is determined by its distance to or deviation from the true population parameter. If, for example, a survey aims to determine the average household income in a certain population, any deviation of the sample estimate from the true value—that is, what would have been determined if all members of the target population were asked their income and they all answered accurately—would decrease accuracy.

Representation and Measurement

There are two types of survey error that harm the accuracy of a survey estimate: random error and systematic error. Whereas random errors are assumed to cancel out each other—that is, negative deviations of the measurement from the true value are compensated by an "equal" amount of positive deviations—systematic errors shift the sample estimate systematically away from the true value; for example, because of certain question wording, respondents in a survey may tend to report a higher number of doctor visits than actually occurred in a given reference period. For linear estimates (such as means, percentages, and population totals), an increase in the random error leads to an increase in variance, whereas a rise in any systematic error results in an ascended bias of the survey estimate. The accuracy of a survey estimate is affected by either an increase of the bias or by a rise of the variance.

In a traditional view, the driving factors or sources of those survey errors are differentiated into two groups: sampling error and nonsampling error. Nonsampling error would then be further differentiated into coverage error, nonresponse error, and measurement error—some older textbooks mention processing error as well. However, a more modern theory-driven approach differentiates *observational errors* and *nonobservational errors*. While observational errors are related to the measurement of a particular variable for a particular sample unit, nonobservational errors occur when a net sample is established that is supposed to represent the target population. Following this path, Robert M. Groves and his colleagues have grouped the sources of error into two primary classes: *representation* and *measurement*:

1. The first class of error sources applies to the representation of the target population by the weighted net sample (representation).
2. The second class of error components adds to the total error by affecting the edited survey responses obtained from a respondent (measurement).

This extension of the traditional total survey error concept provides room for a detailed analysis of the mechanisms by considering several sources of error at the same time, including possible interaction effects between the sources.

Total Survey Error Components Affecting Representation

Coverage Error

For a sample to be drawn, a sampling frame is necessary in order to provide the researcher with access to the members of the population from whom data are to be gathered. The incompleteness of this frame and the possible bias of its composition cause misrepresentations of the population by the sample. If a group is underrepresented in the frame—for example, individuals who own mobile phones as their only telecommunications device are missing from traditional random-digit dialing (RDD) sampling frames because they do not have a landline telephone number—the sociodemographic or substantive characteristics of this group cannot be considered when computing the survey statistic. This causes a lack of accuracy of the survey estimate since some groups might be underrepresented in the frame and subsequently in any sample that is selected from the frame, resulting in coverage bias.

Sampling Error

Once a frame is available, a random sample needs to drawn: for example, a simple random sample, a stratified sample, a cluster sample, or a more complex sample design. Based on this sample, the standard error is computed by taking the square root of the division of the variance in the sample and the number of cases in the sample. The standard error is then used to compute the confidence limits and the margin of error—both are indicators for the precision of the estimate. Accordingly, the magnitude of the sampling error is one key component of the total survey error. It depends heavily on the design of the sample. For a fixed number of sample cases, the standard error usually decreases if stratification is applied. By contrast, a clustered sample is generally characterized by a larger variance which, in turn, raises the sampling error for a particular estimate. However, within a fixed budget, clustering usually increases precision, because the effective sample size can be increased even though the variance suffers from the inflationary design effect (i.e., *deff*) caused by clustering.

Nonresponse Error

Unit nonresponse error is the facet that is the best studied among all bias components within the TSE framework. Since the early days of survey methodology, researchers have been aware of the fact that some portions of the gross sample cannot be reached in the field phase of a survey or are not willing to comply with the survey request for cooperation. Because the responses of those groups may differ considerably from the responses of those members of the gross sample who can be reached and who are willing to cooperate, unit nonresponse is considered a serious source of systematic error that generates nonresponse bias. The literature provides comprehensive theoretical approaches to explain the various stages of respondent cooperation and also findings that can be generalized beyond particular surveys. In part, this is due to the fact that a potential nonresponse bias can be assessed for variables where parameters are available from official statistics (e.g., household income). Compared to other sources of error, this leaves survey researchers in a comfortable situation, as a possible bias can be observed more easily and taken into consideration when survey findings are interpreted.

Adjustment Error

Finally, the net sample needs to be adjusted for design effects introduced by the sample design. If the sample design, for example, would require a disproportional stratified sample, an appropriate weighting procedure would have to be devised to compensate for the unequal selection probabilities when estimating the population parameter. In addition, and as noted earlier in this entry, the net sample may need to be adjusted for possible nonresponse bias. Both procedures require complex computations that take into account information from the gross sample, official statistics, or both. Whereas the first approach may potentially increase the random error of the estimate, the second approach may introduce systematic errors into the sample and thus bias the estimate.

Total Survey Error Components Affecting Measurement

The four sources of error discussed so far were related to the representation of the target population by the weighted net sample. Coverage error, sampling error, nonresponse error, and adjustment error all potentially contributed to the random error or systematic error of the survey estimate. The next three sources of error—specification error, measurement error, and processing error—are concerned with the measurement process.

Specification Error

Most concepts of interest in surveys cannot be observed directly. Instead, the measurement process requires researchers to operationalize and translate the construct into questionnaire items that can be asked by interviewers and answered by respondents. For example, the general public's attitudes about illegal immigration ideally should be decomposed into several questionnaire items about the various dimensions of illegal immigration. Respondents then would be asked to report attitudes on each of these items. The combined score of all items would then be treated as a scaled measurement of the attitudes toward illegal immigration. If an important aspect of this construct were omitted from the scale, then the validity of the operationalization of the construct would be harmed, because the scale would not measure the construct completely and a specification error would occur. This can result in a serious bias, because the estimates based on an incomplete scale often would not mirror the complete true attitudes of the members of the target population on illegal immigration.

Measurement Error

Measurement error is a complex component of total survey error. It consists of various elements that individually and jointly may cause systematic survey error as well as random survey error. Accordingly, measurement error potentially contributes to an increase of the estimate's variance as well as to its bias. Measurement error arises from the mode of survey administration, from the questionnaire or survey instrument and from the setting in which the instrument is administered, from the interviewers (if present), and also from the respondents.

Survey Mode

A traditional trichotomy of data collection modes differentiates face-to-face surveys, telephone surveys, and self-administered (mail and Internet) surveys. They differ with respect to (a) the presence or absence of an interviewer—which allows for various degrees of standardization of the measurement process, for different types of motivational support to the respondent, as well as explanation and help for the respondent—and (b) the dominant communicative channel (audio-visual, audio-only, visual-only). In recent years, many new survey modes have evolved with the introduction of modern information and communication technologies. Some of these modes transfer an established methodology into a computer-assisted mode (e.g., the shift from paper-and-pencil personal interviewing [PAPI] to computer-assisted personal interviewing [CAPI] or computer-assisted telephone interviewing [CATI]), other new modes evolve as a consequence of merging survey modes (e.g., mobile Web surveys that use messenger systems or agents or avatars). Each of these survey modes has its particular strengths and weaknesses for specific survey topics and survey designs. Whereas a Web survey might increase the variance of an estimate because respondents tend to answer a frequency question superficially compared to a face-to-face interview, the response to a face-to-face version of the very same questions might be prone to a higher degree of systematic social desirability distortion because of the presence of an interviewer, which in turn contributes to measurement bias.

Questionnaire

During the past 25 years, questionnaire design has been seriously developed from an art of asking questions to the science of asking questions. This line of research has demonstrated on innumerable occasions that slight modifications in the wording of a question and/or the response categories, or of the order of the questions and/or response categories, or in the visual design of the whole questionnaire, as well as of single questions, can affect the answers obtained from the respondents. Since the early days of the Cognitive Aspects of Survey Measurement (CASM) movement, numerous research papers and textbooks have contributed to a coherent theoretical approach that helps explain and predict random measurement error and systematic measurement error related to the questionnaire.

Respondent

Also within the CASM framework, a detailed theoretical approach on how respondents consider and answer survey questions has been developed. As a result, the question–answer process has been described psychologically in great detail. Using this framework, several systematic and random respondent errors have been identified related to what may happen when respondents answer survey questions. For example, satisficing behavior—as opposed to optimizing response behavior—as well as mood effects have been demonstrated to occur by methodological research.

Interviewer

Finally, it has been demonstrated that the personal and social characteristics of interviewers, if they are present, as well as their task-related and non-task-related behaviors may have a considerable influence on the answers obtained from the respondents. Accordingly, a great deal has been learned in the past 30 years about how to train and monitor interviewers to reduce the likelihood that their behavior will negatively impact respondents' answers. However, it should be recognized that it is impossible to eliminate all of the effects of individual respondent reactions to the personal and social characteristics of an interviewer, as interviewer-administered surveys require a personal encounter of respondents and interviewers.

Processing Error

In addition to lack of specification validity and to measurement error, the errors that may occur when editing and processing survey responses obtained from the respondents are part of the TSE framework. Poor handwriting with open-ended questions, the treatment of answers that were initially not codable, and the classification of occupations are just a few examples of possible errors that may occur at the data-editing stage of a survey. Also, scanning paper forms using OCR (optical character recognition) technology or keying the answers from a paper questionnaire into a database are prone to errors. In addition, crucial responses may need to be imputed as a result of item nonresponse (i.e., missing data), and this is susceptible to random error and to systematic error. Accordingly, these survey steps and the errors associated with them might either increase the variance of a variable—which in turn inflates the standard error and the margin of error—or compromise the accuracy of a response because a bias is introduced.

TSE and a Simplified Formula for Mean Square Error

Statistically speaking, TSE is the difference of a sample estimate and the respective parameter in the target population. This difference is measured by the mean square error (MSE), which in turn consists of two components: (a) the squared sum of the bias components plus (b) the sum of the variance components. For the mean square error, one needs to combine both bias and variance from all sources to obtain an estimate of the TSE. However, although most sources of possible error contribute to bias and to variance simultaneously, some error sources are predominantly responsible for an increase of either variance or bias. Thus, a simplified formula for the mean square error is as follows:

$$\mathrm{MSE} = (\mathrm{B}_{spec} + \mathrm{B}_{meas} + \mathrm{B}_{proc} + \mathrm{B}_{cov} + \mathrm{B}_{nr})^2 + \mathrm{VAR}_{meas} + \mathrm{VAR}_{samp} + \mathrm{VAR}_{adj},$$

where the terms have the following meaning:

B_{spec} = Specification bias (reduced construct validity)
B_{meas} = Measurement bias
B_{proc} = Processing bias
B_{cov} = Coverage bias
B_{nr} = Nonresponse bias
VAR_{meas} = Measurement variance
VAR_{samp} = Sampling variance
VAR_{adj} = Adjustment variance

Even though it is easy to estimate sampling variance, as explained in every introductory statistics textbook, it is less than trivial to estimate the other types of variance and especially the biases. Thus, the MSE as a measure for the TSE is often only of heuristic value, because the exact value of a particular variance or bias component cannot be computed reliably.

The MSE offers the opportunity to evaluate survey designs and the estimates computed based on a survey design. Thus, when reporting the results of a survey, end-users of the particular survey data can assess the quality of the estimate not only based on sampling error and the margin of error but also based on other error components. This is especially important because the bias component of the MSE generally is assumed to exceed the size of the variable error. Thus, the sample estimate of the population parameter often departs more from the true value than what is assumed based on the sampling error alone.

Also, the MSE allows an assessment of various survey designs to facilitate the decision of which design likely would produce data of the highest quality in a given time frame and for a fixed amount of money. However, in practice, survey designs are not only evaluated in terms of their MSE. Instead, survey design A may be preferred even though it produces data of lower quality in terms of the MSE compared to survey design B. For example, if the estimated cost for survey design B is considerably higher that design A's costs, the person responsible for the survey may have no choice but to go with survey design A. Thus, the TSE framework also relates to survey costs and requires survey designers to consider the accuracy in relation to cost and the timeliness of reporting.

Ultimately, the researcher's goal is to reduce the TSE by balancing various trade-offs in design decisions. Most of the time, design decisions—like choosing a certain mode of administration or choosing a special interviewer training procedure—affects not only one source of error but rather multiple sources. Thus, each desirable reduction in terms of a particular error source may be accompanied by an undesirable increase of some other error. Therefore, survey designers need to be able to compromise and balance several sources of error simultaneously.

Limitations of the TSE Framework

Even though TSE offers a convincing framework for the accuracy of a survey estimate, it also suffers from a serious drawback. Currently, the effort necessary to compute a reasonable quantitative estimate of the magnitude for a particular error component usually exceeds the available resources. The estimation of the MSE requires multiple repetitions of the survey design, which is usually too costly and also not feasible because the target population does not remain unchanged in between the repetitions. Also, for many survey designs some error components are not accessible because of the field procedures applied or legal constraints (e.g., privacy laws prohibit extensive nonresponse follow-up studies in many countries). Also, it should be noted that for the exact computation of the MSE, the population parameter needs to be readily available. Because this is usually not the case, the MSE is seldom explicitly determined in practice. More often only a few key components are estimated, or a survey design is rated along the various components of bias and variance on a scale from "low" to "high." The decision for a particular survey design then is made on the basis of a detailed computation of some of the error of the components and a rough assessment of the magnitude of some of the other error components. This leaves the researcher, as well as the end-user of a survey statistic, in a situation where a qualitative assessment of the magnitude of the total survey error is the best available assessment.

Strengths and Benefits of the TSE Framework

Nevertheless, survey research and survey methodology have benefited greatly from the emerging TSE approach. The TSE framework has helped to make researchers more aware of possible errors in their survey statistics and the implications of these likely errors. For example, if the response rate and the size of the net sample are the only noticeable indicators for a given survey, many likely biases remain undetermined. Thus, the TSE framework motivates a systematic reflection on possible impairments of survey quality. In doing so, it stimulates a professional evaluation of ongoing surveys in terms of data quality and provides a common language and terminology for a critical discussion.

In addition, the framework provides a theoretical explanation for the various types of possible errors (variance and bias) and also for the underlying mechanisms (random error vs. systematic error). Also, it names a wide range of possible sources of threats to data quality. Hence the TSE framework suggests a theoretical approach for further developments of the survey methods beyond traditional approaches (ones that are not working well enough). In addition, it provides measurable indicators in order to evaluate the improvements of these new survey methods.

The TSE framework also has provided a basis for heightened interdisciplinary discourse across the boundaries of traditional disciplines. Surveys have been used for a long time in sociology, psychology, economics, and educational research, but until relatively recently, professionals in these disciplines have not been in close communication with each other. Even though it is too early to state a true integration of the field-specific methodologies, one can say that the survey branches of the subject-specific methodologies have merged, or at least are in the process of integration, based on the TSE framework and the methodological advances it has stimulated.

In an international perspective, the integrated concept of a TSE has contributed to the dissemination of "standardized" quality criteria and a set of methods to meet those criteria. International survey endeavors like the Programme for International Student Assessment, the International Social Survey Program, and the European Social Survey would not be feasible if researchers of diverse cultural and disciplinary backgrounds had not begun to interact and cooperate within a common framework. Even though there are still many national specifics in the design and the administrations of a survey, a minimum degree of concordance in the assessment of the data quality is provided by the TSE framework.

From a constructivist perspective, the TSE framework seems to be naive in one of its fundamental assumptions: Is there really something like a true value? Although one could argue that it is a reporting error if a respondent omits a certain portion of his or her income when asked for the monthly gross income of the household (e.g., the portion that comes from child support/alimony), one might also argue that this survey's definition of income also contributes to a social construction of "income." More traceable, surveys contribute to the shape and guise of public opinion when results of attitude surveys are repeatedly reported by the media and thus function as a reference point for the general public while they form their opinions on various public issues. Even from a less fundamental perspective, it remains questionable whether there is a perfect, faultless way of designing and conducting a survey. Accordingly, the true value is rather a chimera that cannot be measured without intervening with instruments and procedures that are, by themselves, selective and incomprehensive in principle.

However, from an analytic point of view, it definitely makes sense to assume fixed and constant true values at a given point in time. And it remains the principal goal of survey methods that they measure and mirror these parameters in the target population. With the aid of the TSE framework, survey researchers and survey practitioners have the instruments at hand to assess, discuss, and improve the quality of the respective estimates.

Marek Fuchs

See also Bias; Cognitive Aspects of Survey Methodology (CASM); Construct Validity; Coverage Error; Design Effect (*deff*); Interviewer-Related Error; Mean Square Error; Measurement Error; Missing Data; Mode-Related Error; Nonresponse Error; Nonsampling Error; Questionnaire-Related Error; Random Error; Respondent-Related Error; Sampling Error; Sampling Frame; Survey Costs; Systematic Error; True Value; Unit Nonresponse; Variance

Further Readings

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. New York: Wiley.

Groves, R. M. (1989). *Survey error and survey cost*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling errors in surveys*. New York: Wiley.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

Touchtone Data Entry

Touchtone data entry (TDE) is a method used with telephone surveys to enable the respondent to directly enter information using the keypad on his or her phone rather than speaking the information to an interviewer. The technology is the same as that used extensively by financial institutions and customer service contact centers. The applications of TDE for survey research include the following:

- As a cost-saving method where the interviewer is replaced by pre-recorded questions delivered by the computer. This is called IVR/TDE, where IVR refers to interactive voice response, and differs from IVR/ASR (automatic speech recognition) in that the latter uses speech recognition in place of TDE.
- To reduce respondent burden when only a small amount of information is needed. For example, respondents to a quarterly business survey might have the option of returning a short paper form, entering the information via a Web site, or dialing in the information using TDE.
- To increase respondent convenience. TDE systems are usually operational 24 hours a day, 7 days a week.
- To provide additional privacy for sensitive questions within an interviewer-administered survey, as the response will not be overheard by someone else in the same room as the respondent (e.g., a parent overhearing a teenager), and the system can be configured to prevent the interviewer from seeing the information entered.

Because of the limitations of most telephone keypads (only 12 keys, including # and *), TDE is best used only for binary responses (e.g., "enter 1 for Yes, 2 for No"), limited choice sets (e.g., "1 for Yes, 2 for No, 3 for Undecided), and simple numeric entry (e.g., "Please enter your year of birth in four digits").

Other numeric information can be collected via TDE, but additional read-back checks (e.g., "You entered 10 thousand and 43 dollars, is that correct? Press 1 for Yes, 2 for No") are necessary because many phones do not have a visual display whereby the respondent can readily see if he or she made an error.

Drawbacks of TDE include the following:

- The need for the respondent to have a touchtone phone
- The increasing tendency of phones to have the keypad embedded in the handset (such as with cordless and mobile phones), making TDE physically awkward as the respondent juggles trying to listen with trying to respond

- Very limited options regarding text, as opposed to numeric, entry due to the many-to-one relationship between alpha characters to numeric keys on keypads
- The need to keep the TDE component of a survey very short and simple (Few people can recall more than four menu items, and if that is combined with a lengthy question administered by a mechanical voice that cannot respond to respondent queries, the respondent may become frustrated and give up.)
- The risk of lower response rates, if a live interviewer who can keep the respondent engaged and motivated is not present

Jenny Kelly

See also Interactive Voice Response (IVR); Telephone Surveys

TRACKING POLLS

A tracking poll is a series of individual surveys repeated continuously over time to measure attitudinal and behavioral changes in a target population. While most commonly associated with election campaigns, tracking polls also are used for a variety of other research needs, from measurement of consumer sentiment to ad tracking in marketing research.

The term *tracking* often mistakenly is used to describe any trend data obtained over time. In fact it correctly applies only to continuous measurements in stand-alone samples. In all applications, the aim of a tracking poll is to produce an ongoing, rather than a one-time or episodic, assessment of evolving attitudes.

Tracking polls provide substantial flexibility in data analysis. The consistent methodology produces useful time trends. Tracking data can be segregated before and after an event of interest—a major policy address, a campaign gaffe, or an advertising launch—to assess that event's impact on attitudes or behavior. These same data can be aggregated to maximize sample sizes for greater analytical power. And tracking surveys can be reported in *rolling averages*, adding new waves of data while dropping old ones to smooth short-term or trendless variability or sampling noise.

The most publicized use of tracking polls is in election campaigns, particularly in the often-frenzied closing days of a contest when campaign advertising spikes, candidates voice their final appeals, voter interest peaks, and tentative preferences become final choices. Campaigns conduct their own private tracking polls to find their best prospects, target their message, and gauge their progress. The news media use tracking polls to understand and report the sentiments behind these crystallizing choices and to evaluate preferences among population groups, as well as to track those preferences themselves.

Election tracking surveys customarily are composed of a series of stand-alone, 1-night surveys, combined and reported in 3- or 4-day rolling averages. In well-designed election tracking polls, the limitations of 1-night sampling on dialing and callback regimens are mitigated by sample adjustments, such as a mix of new and previously dialed sample and the integration of scheduled callbacks into fieldwork protocols.

Election tracking polls have acquired a negative reputation in some quarters because of their reputed volatility. Thoughtful analysis by Robert Erikson and his colleagues, however, has established that volatility in pre-election polls, where it exists, is introduced by idiosyncratic likely voter modeling rather than by tracking poll methodology.

Humphrey Taylor, chairman of the *Harris Poll*, is credited with creating the first daily political tracking polls, a series of twice-daily, face-to-face surveys done for the Conservative Party in the last 4 weeks of the 1970 British general election: One survey tracked media exposure, while the other tracked issue priorities and party preferences on the issues. Taylor (personal communication, January 29, 2007) relates that his tracking data picked up "a collapse in confidence in the Labor government's ability to manage the economy following bad economic news three days before the election. And this was the reason why almost all the polls which stopped polling too soon failed to pick up a last-minute swing to the Conservatives." The Tories won by two points.

Given such sensitivity, tracking polls have developed into an election fixture. In the largest media-sponsored tracking poll in the 2004 election cycle, ABC News and *The Washington Post* conducted 21,265 interviews over the final 32 days of the presidential campaign. In a variation on the election tracking theme, an academic poll that focused on political issues, the National Annenberg Election Survey, interviewed 81,422 respondents in continuous interviewing from October 2003 through Election Day 2004.

In another use of tracking, ABC, with the *Post* (and previously with an earlier partner, *Money* magazine), adapted election tracking methodology in creating the ongoing consumer confidence survey it has conducted weekly since late 1985. Rather than daily sampling, 250 respondents are interviewed each week, with results combined in 4-week rolling averages. By its 21st anniversary in December 2006, the survey comprised more than 280,000 individual interviews.

Despite the attention focused on election tracking polls, by far the heaviest use of tracking surveys is in market and other commercial research, where tracking polls are used to measure consumer behavior or brand attitudes across time—often to gauge customer satisfaction and loyalty, measure advertising exposure and effectiveness, and direct and assess brand image and crisis management efforts.

Gary Langer

See also Election Polls; Media Polls; Pre-Election Polls; Rolling Averages

Further Readings

Erikson, R. S., Panagopoulos, C., & Wlezien, C. (2004). Likely (and unlikely) voters and the assessment of campaign dynamics. *Public Opinion Quarterly*, *68*, 588–601.

Training Packet

A training packet contains the materials that are used to train survey interviewers, either in general interviewing skills or for a specific survey project. The interviewer training packet is an important part of interviewer training and quality data collection. Interviewers often will use the training packet they receive throughout a study to review procedures, and thus it serves an important role in continuous, on-the-job training.

The entire research process relies upon well-trained, knowledgeable interviewers to ensure the success of a project and the integrity of the data. Managers who are tasked with training interviewers must prepare a thorough training curriculum and system to provide the interviewers with the necessary tools to carry out their duties in an ethical and quality-minded manner. Trainers must always keep in mind that many interviewers do not have formal academic experience with the research process. Thus, it is important that interviewer training not only discusses the techniques of interviewing but also the scientific objectives and ethical obligations of researchers. It should be the goal of every interviewing facility and field staff to standardize the interviewing process, which can be quite difficult considering the complexities of social interaction.

The following sections discuss the basic principles necessary for a thorough interviewer training packet and the content that is used to accomplish this.

Administrative

Typically, general training begins with an overview of the research organization and applicable administrative tasks. Administrative information includes time reporting, assignment logging, and work hour scheduling. The addition of an overview of the research firm provides interviewers with the "big picture" by showing them how they fit into the organization as well as the research process.

The Science of Social Research

The science of social research has come a long way in the past half century. It is imperative that interviewers have an understanding of research ethics and the scientific approach. Most interviewers will not be well versed in these ideas.

The research process begins with project development—that is, what and who are the researchers studying. Depending on the research firm, this could come from an outside client or from within the research organization. The second phase includes the development of the sampling methodology and the instrument (questionnaire) design. The third phase is data collection, which could be computer-assisted telephone interviewing (CATI), computer-assisted personal interviewing (CAPI), Web, or some other collection method. Data entry may also be included in this phase depending on the chosen data collection method. The final phase includes data reduction and analysis as well as reporting.

Interviewing

The following outlines important interviewing training topics that should be addressed in the training packet. It is important to note, some of these topics may only be relevant to certain types of interviewing modes, such as CATI or face-to-face (field) interviewing, but most are universally applicable.

Interviewer Bias

It is important to begin interviewer training with a discussion of interviewer bias. Interviewer bias is the altering of respondent answers due to interviewer actions. Interviewers should keep in mind that everything they say or do can affect respondents. They can lead respondents to answer a certain way or be less truthful. Interviewers should never give respondents any inclination as to what their own opinions may be; instead, they should maintain neutrality at all times.

Approach

The approach an interviewer uses to successfully interview a respondent includes the concepts of rapport, interviewer neutrality, and reading questions exactly as they are written. In addition the interviewer training packet should contain contact information and instructions about each of these matters.

Developing rapport is a great asset to an interviewer. It can help prevent a respondent from terminating before finishing the questionnaire, smooth out the interviewing process, and leave the respondent feeling good about participating in a survey. Building rapport begins with the use of a professional but pleasant demeanor and tone. Interviewers must balance being too happy and enthused, which appears very sales-like or "false," with being too scientific, which comes across as cold.

At the same time, interviewers must always take a neutral demeanor when it comes to anything that could distort the answers given by a respondent. They cannot impart any of their own opinions, as this may bias the respondent. There will be times when interviewers will undoubtedly come across respondents with radically different views from their own; however, they must not show surprise or disagreement with the respondent or judge the respondent in any way. Instead, interviewers must always remember that they are there to collect accurate information, not to impart their own judgments upon respondents.

Interviewers are to read the questions exactly as they are written. Interviewers should be informed that the researchers deliberately, and with great effort, analyze each and every word during the development of the questionnaire. Interviewers are not to provide definitions or clarifications for questions unless instructional materials allow for such actions.

Initial Interaction

The first few seconds of contact between interviewer and respondent often set the tone for all that follows. Within the first few seconds, respondents often will have made up their mind whether they will continue speaking with an interviewer or will terminate the contact. Interviewers must build rapport quickly, deliver their introduction, and screen for eligible respondents within this initial interaction.

The training packet should include instructions and examples about how the interviewer can deliver the introduction of the interview with confidence. Interviewers may have some input into the development of the introductory text, but once it is set they cannot alter it beyond their own ability to properly tailor it to each individual respondent.

The introduction is typically followed by a screening section. This should be explained in the training packet and may be complex with multiple questions, or it may be simple with only a single question. Interviewers must be trained to take extra precaution when asking screening questions. All too often, a respondent will automatically answer in agreement with an interviewer only to have the interviewer discover later that the respondent actually did not qualify to participate. Interviewers should be trained to always encourage respondents to think before they respond. This can be done through the instructions about the use of an appropriate pace and thought-provoking tone of voice.

Techniques for Question Delivery

The training packet also should include information about the following techniques used by interviewers to properly administer the questionnaire.

When a respondent gives an ambiguous answer, an interviewer must probe respondents for clarification. Probing is the use of a nonleading phrase to encourage a respondent to provide a more precise response. Probing should be described in the training packet, and examples of appropriate and inappropriate probes should be included. Interviewers must be comfortable with this technique, as it is needed in a variety of situations. It can be used when respondents do not adhere to a closed-ended question or when an unclear open-ended response is given.

Respondents will sometimes stray from the questions. Interviewers must know how to refocus respondents' attention to the task. Focusing is necessary when a respondent becomes irritated by a particular question and proceeds to rant about the failures of the questionnaire or when a respondent gives open-ended responses to closed-end questions. Interviewers can use reassuring phrases demonstrating that they are listening to the respondent but that they must proceed with the interview. The training packet also should include instructions on how to accomplish this.

Jeff Toor

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); Ethical Principles; Face-to-Face Interviewing; Interviewer-Related Error; Interviewer Training; Interviewer Neutrality; Mode of Data Collection; Probing; Respondent-Interviewer Rapport; Tailoring

Further Readings

Barber, A. (1997). Essential interviewing techniques. *National Network of State Polls Newsletter*, *27*, 1–5.

Fowler, F., & Mangione, T. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: Sage.

Gwartney, P. (2007). *The handbook of training and practice in infant and preschool mental health*. San Francisco: Jossey-Bass.

Trend Analysis

Trend analysis is a statistical procedure performed to evaluate hypothesized linear and nonlinear relationships between two *quantitative* variables. Typically, it is implemented either as an analysis of variance (ANOVA) for quantitative variables or as a regression analysis. It is commonly used in situations when data have been collected over time or at different levels of a variable; especially when a single independent variable, or *factor*, has been manipulated to observe its effects on a dependent variable, or *response variable* (such as in experimental studies). In particular, the means of a dependent variable are observed across conditions, levels, or points of the manipulated independent variable to statistically determine the form, shape, or *trend* of such relationship.

Examples of a quantitative variable that a survey researcher may be interested in manipulating to measure its effects on another quantitative variable are amount of incentives provided in a survey (e.g., 0, 2, 4, and 6 dollars), interviewer training time (e.g., 0, 1, 2, 3 hours), number of interviews assigned to interviewers, time allowed to perform a memory task, distance of exit polling interviewers from the voting booth, number of callbacks in a telephone survey, time elapsed between the first occasion on which participants were sent questionnaires and follow-up surveys, among others.

Using trend analysis, the researcher evaluates statistically whether the relationship between the dependent and independent variable is linear, quadratic, cubic, or other high-order function. The number of *bends* to be tested in a polynomial function is determined by the number of *conditions*, or *levels*, of the independent variable. For instance, if there are two conditions in an experiment, only a linear trend can be tested; if there are three conditions, only a quadratic trend is testable; if there are four conditions, a cubic trend is possible, and so on.

To implement this analysis, it is common practice to have an ordered and equally spaced metric for levels of the independent variable and an equal number of subjects allocated exclusively to each condition of such variable. In cases where the number of subjects allocated into each condition varies, weighting procedures may be used; however, there may be limitations associated with weighted trend analysis that the researcher should further identify. Additionally, interpretation of trend analysis results requires attention to surrounding aspects such as statistical power and effect size.

If implemented as regression analysis, the independent variable (X) is entered into the equation and followed sequentially by increasing powers of the same variable. Consequently, regression equations for trend analysis may be specified as $Y = a_1X + b$ (linear), $Y = a_1X + a_2X^2 + b$ (quadratic), $Y = a_1X + a_2X^2 + a_3X^3 + b$ (cubic), and so forth, where Y represents the dependent variable, b the intercept, and $a_1 \ldots a_k$ regression coefficients. The highest-order regression coefficient in the equation being statistically significant settles the shape of the relationship. p-values help determine whether the observed trend is due to a systematic influence of the manipulated variable or by chance alone.

Under ANOVA, trend analysis is conducted using *contrast coefficients*, also known as *polynomial* *contrasts*. These coefficients are numbers usually taken from either available tables or user-designed tables. They help represent a hypothesized trend. For example, contrast coefficients for testing a linear trend would be –1.5, –0.5, 0.5, and 1.5, if the independent variable had four levels. For a quadratic trend they would be 1, –1, –1, 1; for a cubic, –1, 3, –3, 1. Contrast coefficients can be plotted over the *y*-axis across levels of the independent variable to visualize the hypothesized trend, for example, a straight line, a u-shape line, an s-shape line, or higher-order trends. Departures of the actual trend (plot of actual means) from the hypothesized trend (plot of contrast coefficients) are analyzed in terms of *p*-values to determine the type of relationship. Currently, commercial statistical packages can perform trend analysis, but they require user-entered contrast coefficients.

René Bautista

See also Analysis of Variance (ANOVA); Dependent Variable; Experimental Design; Independent Variable; *p*-Value; Regression Analysis; Sample Size; Statistical Power

TRIAL HEAT QUESTION

Nearly all pre-election polls include some type of measure of voter choice in one or more candidate races. Unlike many other kinds of survey measures, voter choice is one of the few in which a definitive and highly visible external validation exists. Polls conducted near the time of an election can be compared with the actual election results.

The key measure used in election polling is the *trial heat question:* the question that asks a respondent how he or she is going to vote in a given race. Although such questions might appear to be very straightforward, there are numerous choices in deciding exactly how the question should be worded and where it should be placed in the questionnaire.

Wording of Trial Heat Questions

Among the choices to be made are (a) whether to include the names of all candidates who are on the ballot or just the major candidates, (b) whether to mention the party affiliation of each candidate named, (c) whether to probe respondents for a choice if they say they are undecided, and (d) whether to obtain some measure of strength of support or certainty of the choice.

Nearly all trial heat questions ask respondents to react as if "*the election were held today*." Most do mention the party affiliation of candidates, though party labels are not on the ballot in all states. There are many ways to probe certainty of choice or strength of support. Most pose a binary choice: strong supporter or not strong, certain or not. One question used by some organizations asks about the candidate not chosen: *Do you think there is a chance that you might vote for _____ in November, or have you definitely decided not to vote for (her)(him)?*

The decision regarding whether or not to include third-party and minor-party candidates is often not a straightforward one. In 2000 and in 2004, Ralph Nader had no realistic chance of winning the presidency, but his candidacy did attract enough support to have arguably affected the outcome of the race in one or more states. All major polling organizations in both years included Nader and his running mate in poll questions conducted in states where Nader was on the ballot. Other third-party candidates were qualified for the ballots in some states, but most polls did not include their names in the question wording asked of respondents, although respondents could volunteer their names.

One practice common to nearly all election pollsters is the rotation of the order of presentation of candidates in the question. Experiments have shown that the order can create small biases in response, and rotation is designed to prevent this. An alternative view is that candidates should be presented in the order they will appear on the ballot (which is itself often determined by chance but then fixed for all ballots in a given jurisdiction). It is not clear whether preserving the ballot order in a telephone survey, for example, re-creates the same cognitive experience a voter will have when encountering the printed ballot or a voting machine on Election Day. In any event, it is almost never feasible for national surveys to implement a state-by-state ballot order.

Pollsters face a somewhat more difficult decision in primary elections, where the number of candidates may be large and no simple criterion, such as the likelihood of winning, can be used to exclude some of the candidates. In presidential campaigns where there is no incumbent, it is not unusual for the field of serious contenders (at least as assessed by political position or experience) to include 10 or more candidates. Until the primary and caucus process gets under way and the field begins to be winnowed, it may be difficult to find grounds for excluding candidates. Thus, pre-primary telephone polls in such instances may require that respondents be read lengthy lists of candidates.

Generic Ballot Questions

A different type of trial heat question does not match the candidates against one another by name but rather asks about voter intention to vote for one party or the other in the election. So-called *generic ballot questions* are commonly used by national polling organizations trying to gauge voter sentiment in elections for the U.S. House of Representatives, where it is not feasible to substitute the names of actual candidates in all of the districts covered by a typical national telephone survey. Generic ballot questions in U.S. House elections have been generally accurate in forecasting the party division of the national House vote in nonpresidential years, despite the fact that only a small percentage of House races are truly competitive in any given election.

Placement of Trial Heat Questions

There is a consensus that trial heat questions should be placed near the front of pre-election polls, with few if any politically substantive questions coming before them. The reason for this is that questions about issues or candidate qualities or traits can raise particular considerations that could affect the choice expressed by respondents, especially among those who are undecided or only weakly committed to a candidate. Voters on Election Day get no such "warm-up" before casting a ballot.

Dealing With Undecided Voters

In any election poll, some respondents will decline to express a choice in response to a trial heat question. It is standard practice to ask these respondents if they "lean" toward one of the candidates. Most polls include the "leaners" in the percentage supporting each candidate.

Even after being asked which way they lean, a small percentage of undecided voters will remain undecided (noncommittal). Pollsters have different ways of dealing with this group. A default assumption is that undecided voters will split in the same proportions as decided voters. Simply omitting the undecided voters achieves the same result. Others split the undecided half-and-half. But most attempt to make some type of *allocation of undecided voters*, based on other responses they have given in the survey. This allocation may be based on party affiliation, images of the candidates, or even responses to a further probe for a candidate preference (known as the "whiny leaner" question because the interviewer typically tells the respondent that "my supervisor insists that I put down something for your preference so could you please help me out?"). Nonetheless, many of those who decline to state a preference after the standard leaner question has been asked may be unlikely to vote, lacking a strong commitment to one of the candidates.

Other questions ask for the certainty of support, strength of support, and even whether the respondent would ever be willing or unwilling to vote for the candidate.

Scott Keeter

See also Election Polls; Horse Race Journalism; Leaning Voters; Undecided Voters

Further Readings

Daves, R. P. (2000). Who will vote? Allocating likelihood to vote and modeling a probable electorate in preelection polls. In P. J. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 206–223). Washington, DC: CQ Press.

Daves, R. P., & Warden, S. P. (1995). Methods of allocating undecided respondents to candidate choices in pre-election polls. In P. J. Lavrakas, M. W. Traugott, & P. V. Miller (Eds.), *Presidential polls and the news media* (pp. 101–123). Boulder, CO: Westview Press.

Visser, P., Krosnick, J., Marquette, J., & Curtin, M. (2000). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. J. Lavrakas & M. W. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 224–262). Washington, DC: CQ Press.

Troldahl-Carter-Bryant Respondent Selection Method

Researchers often desire a probability method of selecting respondents within households after drawing a probability sample of households. Ideally, they wish to simultaneously maximize response rates (by gaining the respondent's cooperation) and within-unit coverage (by obtaining an accurate listing of all eligible persons in a household). However, they need to balance these two goals because no selection technique is perfect at accomplishing both. In the early history of telephone surveys, one procedure that was considered easy, quick, and likely to improve respondent cooperation was known as the Troldahl-Carter method, to which Barbara Bryant later suggested some modifications. Face-to-face surveys commonly used the Kish technique, which asks for a listing of all men and all women in the household, ranked by age.

Verling Troldahl and Roy Carter feared that informants (those who answer the phone) in a telephone survey would become suspicious of questions that try to obtain a listing of all eligible household members and refuse to participate. So they suggested a brief procedure, building on Kish's work but requiring only two questions: (1) *How many persons (18 years or older) live in your household . . . counting yourself?* and (2) *How many of them are men?* Once interviewers knew the basic household composition by sex and age, according to responses to the two questions, they requested the appropriate respondent after consulting one of four versions of simplified selection tables assigned to households randomly. Because this selection method specified that the respondent be the youngest or oldest male or female, it involved a small violation of random sampling and full coverage because some adults in households of three or more adults of the same sex had no opportunity to be chosen. The amount of bias depends on the proportion of persons in the population barred from the sample and the degree to which they differ from the respondents in the variables studied. In addition, in three-adult households, one of the adults would have the chance of selection into the sample twice. Troldahl and Carter thought these biases were minimal. (Kish's plan also contained a small violation of random sampling, although it is considered to be an almost pure probability method.) In subsequent testing of the Troldahl-Carter method, other modifications to it were suggested, such as changing the second question to *How many of them are women?*

Troldahl and Carter provided an example of an interviewer's selection sheet. The potential numbers of adults in the household were shown at the top (1, 2, 3, or 4 or more). Interviewers circled the correct number and drew a line down the column under that number to the number of men in the household, ranging from 0 to 4 or more. For example, if there were three adults in the household and two were men, one version of the tables showed that the column at which those items intersected designated the respondent as "youngest man." The interviewers' instructions were to say, *I have to ask some questions of the* [PERSON SELECTED] *in your household.* If the informant was of the correct gender, the interviewer asked, *Would that be you?* If the desired respondent was of the opposite sex, the interviewer asked for the chosen person and implemented follow-up procedures if he or she was unavailable.

By the mid-1970s, the distribution of women and men within households had changed, although the proportion of men to women had not. Young males were less likely to be at home than they had been in the 1960s. Thus, Bryant suggested a solution to the problem of obtaining better representation of males and females, though she had no suggestions about the problem of unavailable young males. She proposed using the fourth Troldahl-Carter table just half as often as the other three. The sequence of tables used was 1, 2, 3, 4, 1, 2, 3. Many researchers adopted this idea, and this modification became known as the Troldahl-Carter-Bryant (T-C-B) respondent selection method. Ron Czaja, Johnnie Blair, and Jutta Sebestik compared the T-C-B method with a further modification of T-C-B asking for women instead of men, as well as with the Kish procedure. In T-C-B/women, the order of tables is 1, 2, 3, 4, 2, 3, 4. They concluded that the T-C-B/men sample had the highest proportion of males, followed by the T-C-B/women sample. The T-C-B/women version had the best completion rate. The T-C-B/men adaptation produced the highest proportion of married-person households, as well as a lower proportion of one-adult households, possibly linked to sex differences and a lower cooperation rate. Later work has attempted to simplify the T-C-B method, such as the Hagan and Collier technique or the "youngest male/oldest female" method.

Cecilie Gaziano

See also Hagan and Collier Selection Method; Kish Selection Method; Within-Unit Coverage; Within-Unit Selection

Further Readings

Bryant, B. E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research*, *12*, 129–135.

Czaja, R., Blair, J., & Sebestik, J. P. (1982). Respondent selection in a telephone survey: A comparison of three techniques. *Journal of Marketing Research*, *19*, 381–385.
Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.
Hagan, D. E., & Collier, C. M. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly*, *47*, 547–556.
Keeter, S., & Fisher, K. (1997–1998). Comparison of two respondent selection methods. *National Network of State Polls (NNSP) Newsletter*, *31*(Winter), 1–3.
Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*, 380–387.
Troldahl, V., & Carter, R. E., Jr. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, *1*, 71–76.

True Value

A true value, also called a *true score*, is a psychometric concept that refers to the measure that would have been observed on a construct were there not any error involved in its measurement. Symbolically, this often is represented by $X = T + e$, where T is the (error-free) true value, e is the error in measurement, and X is the observed score. For example, a survey respondent's true value (T) on a measure (e.g., a 5-item Likert scale measuring attitudes toward her ancestor's native land) might be 13, but her observed score (X) on a given day may be 15 or 17 or 10. There are many reasons why the observed score may deviate from the true value, but the researcher will rarely know exactly why or exactly by how much. In theory, if the error on the observed score is truly random, then a researcher could take the same measure from the same respondent over and over again, and assuming the true value on that measure for that respondent did not actually change over the time the various measures were being taken, then the mean (average) score of the various observed scores would be the true value. The concept of a true value relates to the concepts of reliability and validity.

For survey researchers, the true value has some very practical applications. First, it reminds researchers that random error is always to be assumed and that any one measurement should never be regarded as error-free. Second, if the researcher has the test–retest reliability on a given variable (x), then the researcher can apply a *correction for attenuation* to adjust any correlations that are calculated between x and some other variable of interest. The correction of attenuation recognizes that any observed correlation between two variables will be suppressed (i.e., lower than their true correlation) by a function linked to the amount of unreliability in the observed scores for the variables. That is, only when there is no error in the observed variables, and thus the observed scores are in fact the true values, will the observed correlation be the true correlation value.

Paul J. Lavrakas

See also Construct Validity; Measurement Error; Random Error; Reliability; Test–Retest Reliability; Validity

Further Readings

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Trust in Government

Trust in government has multiple meanings, many sources, and numerous consequences. David Easton distinguished between *diffuse support*—that is, trust in the institutions and mechanisms of government—and *specific support*, trust in current officeholders. Researchers also can ask about trust in specific institutions within government (legislative, judicial, executive, law enforcement), at all levels (federal, state, and local).

Political life, even at a routine and mundane level, entails risk. When we vote for a candidate, pay our taxes, or obey the law, we hope that our efforts will not be wasted or exploited. Civic life, in a nutshell, demands that we trust our government to uphold its end of the democratic bargain. Survey research has tracked the rise and fall in political trust over the decades and has tried to identify the factors that sustain this tenuous but critical component of civilization.

Many surveys use a single item that simply asks respondents how much they "trust the government." By far the most frequently used item is Question 1 from the biennial National Elections Study (NES): *How much of the time do you think you can trust the government in Washington to do what is right—just* *about always, most of the time, or only some of the time?*

But what exactly is being measured? Many scholars criticize the NES scale (consisting of four items) for tapping predominantly short-term feelings about current office-holders, rather than long-term beliefs about the institutions and mechanisms of government. Because continuity is everything in the study of political trends, however, the NES instrument is likely to remain.

Survey results indicate that trust in government has declined substantially since the 1960s. A majority of the American public these days trusts the government "only some of the time." This has caused great concern for the survival of government and civic life in general. Several scholars have proposed a reciprocal relationship between trust in government and *social capital*, which can be thought of as trust in one's fellow human beings. To investigate this hypothesis, scholars have used data from the General Social Survey, which asks participants to report on the *confidence* they hold in various institutions, including the Executive Branch, the Supreme Court, and Congress. Congress, it turns out, is the institution Americans love to hate more than any other.

Fluctuations in trust follow recent government performance as well as political scandal, but a dislike for today's politicians does not necessarily mean rejection of our way of government. Standard trust measures perhaps exaggerate public disregard for government, for they lump together *skeptical* respondents with the truly *cynical*. When the scale is extended on the less trusting end, only about a quarter of the population place themselves below the midpoint. Still, trust in government has never approached the high point it reached in the 1960s, even after the economic boom of the late 1990s. The accumulation of political scandal, beginning with Watergate and Vietnam, seems to have soured a whole generation on politicians, if not the basic foundation of democratic government.

Polls taken in the months following the September 11 attacks showed an extraordinary jump in trust. This was commonly seen as a "rally round the flag" effect, wherein the attacks inspired extraordinary national unity. A closer look revealed that the attacks merely temporarily shifted respondents' focus to international affairs, where the government has always enjoyed greater public confidence.

Thomas E. Nelson

See also General Social Survey (GSS); National Election Studies (NES); Public Opinion; Social Capital; Trend Analysis; Validity

Further Readings

Hibbing, J. R., & Theiss-Morse, E. (Eds.). (2001). *What is it about government that Americans dislike?* New York: Cambridge University Press.

Langer, G. (2002). Trust in government . . . to do what? *Public Perspective*, (July/August), 7–10.

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annual Review of Political Science*, *3*, 475–508.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, *6*(1), 65–78.

t-Test

A *t*-test is a statistical process to assess the probability that a particular characteristic (the mean) of two populations is different. It is particularly useful when data are available for only a portion of one or both populations (known as a *sample*). In such a case, a *t*-test will enable an estimate of whether any differences in means between the two groups are reflective of differences in the respective populations or are simply due to chance. This statistical process is called a *t*-test because it uses a *t*-distribution to generate the relevant probabilities, typically summarized in a *t*-table.

There are many types of *t*-tests, each of which is appropriate for a particular context and nature of the data to be analyzed. Each *t*-test has its own set of assumptions that should be checked prior to its use. For instance, comparing the average value of a characteristic for independent samples (i.e., when individuals in each sample have no systematic relationship to one another) and for dependent samples (such as when related pairs are divided into two groups) requires different kinds of *t*-tests.

Frequently, a *t*-test will involve assessing whether a sample comes from a specific population or whether average values between two samples indicate differences in the respective populations. In the latter case, three factors determine the results of a *t*-test: the size of the difference in average value between the two groups; the number of data points, or observations, in each group; and the amount of variation in values within each group.

A t-test is conducted by calculating a t-statistic and by using a t-table to interpret its value. The equation for the t-statistic depends on the context and nature of data. For instance, when comparing the likelihood that a sample belongs to a particular population, the equation for the t-test is $t = (x - \mu)/(s/\sqrt{n})$ where x is the mean for the sample, μ is the known value for the population, s is the standard deviation of the sample, and n is the number of data points in the sample.

As an example, in a study that examines the effectiveness of a new math curriculum, researchers might ask whether the curriculum is related to students' state standardized math test scores. A survey might be used to collect state test score data for students who participate in the new curriculum and for those who use a different curriculum. The researchers would want to make generalizations for all students who use and do not use the curriculum. However, because gathering the test score data for every student might be difficult and expensive, the researchers might send the questionnaire to only a sample of students in each group.

After calculating the average test score of each sample, the researchers could use a t-test to estimate the likelihood that the difference between the two samples' average test scores was really reflective of different test scores between the populations and not simply due to chance. If, for instance, the averages of the two samples were very similar, data were only available for a handful of students in each sample, and students' test scores in each sample varied greatly, then a t-test would likely show that the two populations did not necessarily have different average test scores and that the differences in the samples were simply due to chance. This would be shown by a very low value of the t-statistic. If, on the other hand, the difference in average test scores between the samples was great, there were many students in each sample, and students' scores within each sample did not vary greatly, a t-test would support the conclusion that the populations' average test scores were truly different from one another. This would be evidenced by a very high value of the t-statistic.

Joel K. Shapiro

See also *F*-Test; Mean; *n*; *p*-Value; Regression Analysis; Research Hypothesis; Statistic; Type I Error

Further Readings

Kvanli, A. H., Guynes, C. S., & Pavur, R. J. (1986). *Introduction to business statistics*. St. Paul, MN: West Publishing.

Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics* (5th ed.). New York: Wiley.

Type I Error

Type I error refers to one of two kinds of error of inference that could be made during statistical hypothesis testing. The concept was introduced by J. Newman and E. Pearson in 1928 and formalized in 1933. A Type I error occurs when the null hypothesis (H_0), that there is no effect or association, is rejected when it is actually true. A Type I error is often referred to as a *false positive*, which means that the hypothesis test showed an effect or association, when in fact there was none.

In contrast, a Type II error occurs when the null hypothesis fails to be rejected when it is actually false. The relation between the Type I error and Type II error is summarized in Table 1.

The probability of a Type I error is usually denoted by the Greek letter alpha (α) and is often called the *significance level* or the *Type I error rate*. In the table, the Greek letter beta (β) is the probability of a Type II error. In most studies the probability of a Type I error is chosen to be small—for example, 0.1, 0.05, 0.01—or expressed in a percentage (10%, 5%, or 1%) or as odds (1 time in 10, 1 time in 20, or 1 time in 100). Selecting the alpha level of 0.05, for example, means that if the test were to be conducted many times where the null hypothesis is true, one can expect that 5% of the tests will produce an erroneously positive result. A Type I error is an error due to chance and not because of a systematic error such as model misspecification or confounding.

A Type I error could be illustrated with an example of a disease diagnostic test. The null hypothesis is that a person is healthy and does not have a particular disease. If the result of a blood test to screen for the disease is positive, the probability that the person has the disease is high; however, because of the test used, a healthy person may also show a positive test result by chance. Such a false positive result is a Type I error for the disease diagnostic test. Note that a Type

Table 1 Type I and Type II errors in hypothesis testing

	True State of the Null Hypothesis (H_0)	
Statistical Decision	H_0 True	H_0 False
Reject H_0	Type I error (i.e., wrong decision) Probability $= \alpha$	Correct decision Probability $= 1 - \alpha$
Do not reject H_0	Correct decision Probability $= 1 - \beta$	Type II error (i.e., wrong decision) probability $= \beta$

I error depends on the way the hypothesis test is formulated, that is, whether "healthy" or "sick" is taken as the null condition. In survey research, an example of a Type I error would occur when a pollster finds a statistically significant association between age and attitudes toward normalization of U.S. relations with Cuba, if in fact no such relationship exists apart from the particular polling data set.

When testing multiple hypotheses simultaneously (e.g., conducting post-hoc testing or data mining), one needs to consider that the probability of observing a false positive result increases with the number of tests. In particular, a *family-wise* Type I error is making one or more false discoveries, or Type I errors, among all the hypotheses when performing multiple pairwise tests. In this situation, an adjustment for multiple comparison, such as Bonferroni, Tukey, or Scheffe adjustments, is needed. When the number of tests is large, such conservative adjustments would often require prohibitively small individual significance levels. Y. Benjamini and Y. Hochberg proposed that, in such situations, other statistics such as false discovery rates are more appropriate when deciding to reject or accept a particular hypothesis.

Georgiy Bobashev

See also Alternative Hypothesis; Errors of Commission; Null Hypothesis; Statistical Power; Type II Error

Further Readings

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*, 289–300.

Neyman, J., & Pearson, E. S. (1967). On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. In J. Neyman & E. S. Pearson (Eds.), *Joint statistical papers of J. Neyman and E. S. Pearson* (pp. 1–66). London: Cambridge University Press.

Neyman, J., & Pearson, E. S. (1967). The testing of statistical hypotheses in relation to probabilities a priori. In J. Neyman & E. S. Pearson (Eds.), *Joint statistical papers of J. Neyman and E. S. Pearson* (pp. 186–202). London: Cambridge University Press.

Type II Error

Type II error refers to one of two errors that could be made during hypothesis testing. The concept was introduced by J. Newman and E. Pearson in 1928 and formalized in 1933. A Type II error occurs when the null hypothesis (H_0), that there is no effect or association, fails to be rejected when it is actually false. A Type II error is often referred to as a *false negative* because the hypothesis test led to the erroneous conclusion that no effect or association exists, when in fact an effect or association does exist. In contrast to Type II errors, a Type I error occurs when the null hypothesis is rejected when it is actually true. The features of Type II and Type I errors are summarized in Table 1, above.

In the table, the maximum probability of a Type II error is denoted by the Greek letter beta (β), and $1 - \beta$ is often referred to as the *statistical power of the test*. It is intuitive to require that the probability of a Type II error be small; however, a decrease of β causes an increase in the Type I error, denoted by Greek letter alpha (α), for the same sample size.

In many statistical tasks, data collection is limited by expense or feasibility. Thus, the usual strategy is to fix the Type I error rate and collect sufficient data to give adequate power for appropriate alternative hypotheses. Although power considerations vary depending on the purpose of the study, a typical requirement for the power is 0.8, which corresponds to a 20% Type II error rate. However, when data exist in abundance, this strategy will lead to rejecting the null hypothesis in favor of even tiny effects that might have little practical value. For practical use, it is important to have a clear and measurable statement of the alternative hypothesis. For example, if H_0 for a continuous effect x states that there is no effect (e.g., $x = 0$) and the alternative hypothesis (H_1) states that there is some effect (e.g., $x \neq 0$), the concept of Type II error is not practical because H_1 covers all possible outcomes except a single value (that for H_0). Commonly, an alternative hypothesis is given in terms of the measurable effect size, based on the scientific relevance.

For example, the use of mammography to screen for breast cancer provides an illustration of how Type II error operates. The null hypothesis is that a subject is healthy. A positive test result does not necessarily mean that a woman has breast cancer. The main purpose of the mammogram is to not miss the cancer if it is present, that is, to minimize the Type II error. However, tests like mammograms have to be designed to balance the risk of unneeded anxiety caused by a false positive result (Type I error) and the consequences of failing to detect the cancer (Type II error). In survey research, an example of a Type II error would occur when a pollster fails to find a statistically significant association between age and attitudes toward normalization of U.S. relations with Cuba, if in fact such a relationship exists apart from the particular polling data set.

If a number of studies have been conducted on the same topic, it is possible sometimes to pool the data or the results and conduct a meta-analysis to reduce type II error. Because of the relationship between the two types of error, many considerations regarding the Type I error—such as multiple tests and rejecting (H_0) by chance—also apply to the Type II error.

Georgiy Bobashev

See also Alternative Hypothesis; Errors of Omission; Null Hypothesis; Statistical Power; Type I Error

Further Readings

Neyman, J., & Pearson, E. S. (1967). On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. In J. Neyman & E. S. Pearson (Eds.), *Joint statistical papers of J. Neyman and E. S. Pearson* (pp. 1–66). London: Cambridge University Press.

Neyman, J., & Pearson, E. S. (1967). The testing of statistical hypotheses in relation to probabilities a priori. In J. Neyman & E. S. Pearson (Eds.), *Joint statistical papers of J. Neyman and E. S. Pearson* (pp. 186–202). London: Cambridge University Press.

U

UNABLE TO PARTICIPATE

The *unable to participate* survey disposition is used in all types of surveys, regardless of mode. It occurs when the selected respondent for a survey is incapable of completing the telephone or in-person interview or of completing and returning the Web-based or mail questionnaire. Cases coded with the unable to participate disposition often are considered eligible cases in calculating survey response rates.

A variety of permanent and temporary reasons may cause respondents to be unable to participate in a survey. Permanent reasons for being unable to participate in a survey include language barriers, physical or mental disabilities, and chronic severe illnesses. The sampled respondent in a telephone or in-person survey may not speak English (or another target language) well enough to complete the interview. Respondents with a mental disability may not have the cognitive capacity to complete the survey questionnaire or the interview. Being hospitalized beyond the survey's field period would also be considered a permanent reason for being unable to participate in the survey. In the case of a mail or Internet survey, being illiterate would constitute a permanent reason for being unable to participate. The unable to participate survey disposition usually is considered a final disposition in cases where there is a permanent or long-term reason for the respondent not being capable of participating in the survey.

In other cases, the reason for a respondent being unable to participate in a survey may be temporary in nature. For example, the sampled respondent might be intoxicated or might be ill or in the hospital for a brief period. In these cases, the unable to participate disposition should be considered a temporary disposition although the interviewer should do his or her best to determine how long this situation will last (even if just an estimated number of days) and note this in the case history file; then the case should be recontacted after an appropriate amount of time has passed.

It is important to note that cases in which the respondent indicates that he or she is unable to participate in the survey may be a tactic by the respondent to avoid completing the interview or the questionnaire. Although these instances are not common, cases in which it is detected should be considered indirect respondent refusals.

Matthew Courser

See also Final Dispositions; Language Barrier; Respondent Refusal; Response Rates; Temporary Dispositions; Unavailable Respondent

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

UNAIDED RECALL

One of the first decisions a designer of a questionnaire must make is whether the information sought can best and most appropriately be gathered from respondents by asking questions with or without clues to the possible or most appropriate answers. A question-asking strategy that relies on as few cues as possible, or even none whatsoever, in order to encourage the respondent to spontaneously mention items of interest is known as *unaided recall*. Such question tools are commonly used in marketing research and other situations in which it is desirable to know the respondent's reaction to questions unprompted by previous questions or topics mentioned in previous questions. In social science surveys, it is often desirable to elicit answers unstructured by other specific political or social questions. These types of subtle cues and reminders have a way of prompting, structuring, and defining the range of suitable ideas and events to be recalled.

Unaided recall questions typically ask respondents to mention specific instances of more general phenomena that are typically construed rather broadly. For example, an unaided recall question might ask respondents to describe all television news programs they watched in a given day or week. Such a question design places the cognitive burden on the respondent to remember, name, and categorize his or her answer. The strength of such a question is that it allows the respondent to respond naturally, that is, unprompted by various cues as to what kind of specific information is sought. However, this question format can be subject to a variety of errors, including accuracy and omission. On occasion, for certain question topics, accuracy can be assessed by comparing the respondents' unaided responses to various types of records. Omission of useful information is likely when questions are unaided because particular information might easily be forgotten or considered irrelevant by the respondent. Cues, as in *aided recall*, can improve memory for suitable events and ideas by providing examples, but they can also introduce error by suggesting inappropriate ideas to the respondent. Most importantly, however, cues can affect and even structure the kinds of ideas and events that the respondent mentions.

People's memory for particular things may fail for many reasons, including that they may have never really known the information, have honestly forgotten it, or they are unwilling to spend the required time and mental effort to recall completely and make sure their answers are as accurate as can be. People might remember the information incompletely, recall it incorrectly, or express it poorly so that it is misunderstood by the researcher.

However, sometimes researchers want to know exactly whether a respondent can instantaneously remember (recall) something without being prompted with any clues, such as when branding studies are done and respondents are asked on the spur of the moment to come up with a name of a company (e.g., *Please name a company that makes digital video recorders*). In these cases, it is informative to the survey sponsors (e.g., the client company) to know what portion of the sample is able to spontaneously generate their name and their competitors' names without any prompting. In such a case, unaided recall would be the preferred research strategy. This does not preclude use of additional aided recall questions later in the survey.

Certain types of unaided recall question characteristics can affect memory. In designing a survey question to ask people about certain events, the survey designer could allow the respondent to list the relevant events in any order. When faced with particular tasks, however, such as biographical information, it may be most effective to ask in reverse chronological order, depending perhaps on the topic domain. The survey designer would also want to make certain that respondents have sufficient time to think the issue through carefully before responding and not be hurried along by the interviewer. Survey designers can also aid respondent recall by including separate questions about smaller, less inclusive categories of events. A question requiring respondents to recall interactions with a physician might inquire separately about office visits and email or telephone consultations.

Another significant problem of the unaided recall question strategy relates to the covered time frame. One problem is *telescoping*, which is a type of mental bias such that a respondent may recall a specific event but misremember the time frame in which it occurred. As Seymour Sudman and Norman Bradburn have noted, telescoping can be forward, in which people recall events as occurring more recently than they really did, or backward, in which people recall events as happening longer ago than they really did. Telescoping poses significant problems for estimation, as this type of bias may increase as the time period becomes shorter. This can cause large problems of overestimation.

Bounded recall is one question-asking procedure that reduces telescoping bias by repeating interviews with the same respondents over periods of time. The first round of questions is typically unstructured, with repeated rounds involving various structured questions in which the answers to previous unstructured questions are incorporated as reminders. At the same time, new behaviors are checked with previous ones to avoid duplications. In this way, the first round of unaided questions can be said to bound the responses and provide a kind of baseline for subsequent rounds of questions and answers. Bounding can reduce telescoping but does not improve the problem of omissions. Bounding, in its full implementation, requires expensive panel data in which the same respondents are repeatedly interviewed. But modified bounding questions can be implemented in a single interview by asking first about an earlier time period and then referring back to that period when asking about subsequent time frames.

The mode by which the question is administered is also important to consider, that is, face-to-face, telephone, or self-administered interview, such as in a mail or Internet survey. If the questionnaire is self-administered, presumably the respondent could take as much time to think about the answer as needed. A person might feel under more time pressure if an interviewer is waiting for the answer on the telephone than in person. In self-administered online questionnaires, the designer might want to control cueing effects of question context or order by requiring people to answer questions in a fixed order and present them one at a time, thus limiting people's ability to scroll ahead to learn more about the question context. If the question is being posed to a respondent by an in-person interviewer, it also places some burden on the interviewer to gather verbatim answers on the fly as the respondent provides the information. To ensure this information is gathered completely and accurately, the interviewer must be carefully instructed to reproduce the respondent's comments with precision.

Accuracy and completeness can be enhanced by training interviewers to use procedures such as asking the respondent to slow down, to repeat, and to review the answer with the respondent more than once to make sure the answer accurately reflects what the respondent intended. Such interviewers might also be equipped with voice recording technology or be specifically trained to write down material rapidly and accurately. Interviewers must also be given specific and uniform instructions regarding the appropriate probing that they are to do to make certain that the respondent has told all they need to in order to answer the questions completely. Standard probing protocols are important to minimize the likelihood that a respondent's ability to provide answers is not confounded with loquaciousness and perhaps general social skills.

Commonly anticipated answers to unaided recall questions can be precoded and be available for selection by the interviewer, but they are not to be read to the respondent. This strategy requires judgment on the part of the interviewer to recognize and categorize the answer properly. Unaided recall is an important tool in the design of questionnaires.

Gerald M. Kosicki

See also Aided Recall; Aided Recognition; Bounding; Cognitive Aspects of Survey Methodology (CASM); Errors of Omission; Interviewer Training; Mode of Data Collection; Precoded Question; Probing; Satisficing; Telescoping; Verbatim Responses

Further Readings

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Tanur, J. M. (Ed.). (1992). *Questions about questions: Inquiries into the cognitive biases of surveys.* New York: Russell Sage Foundation.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York: Cambridge University Press.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research.* Chicago: University of Chicago Press.

Unavailable Respondent

The *unavailable respondent* survey disposition is used in all types of surveys, regardless of mode. It occurs when the selected respondent for a survey is not available (a) at the time of interview, (b) for some portion of the survey field period (temporarily unavailable), or (c) until after the field period of the survey has ended (permanently unavailable). Cases in which the respondent is unavailable, regardless of the duration, are considered eligible cases in calculating survey response rates.

In most surveys, less than half of the completed interviews result after the first contact with the respondent, regardless of whether that contact occurs in person, by telephone, by mail, or by Internet. Many respondents simply are not available at the time of an interviewer's telephone call or visit to the respondent's residence. Respondents may also be away from their residence, may not have retrieved or processed their postal mail or their electronic mail, or may be too busy to complete the survey immediately after it is received. In these situations, it is important to make additional contacts with the selected respondent. For example, in-person and telephone interviewers may ask questions to determine a good day or time for another call or visit to the respondent; survey firms may send a reminder postcard or email message to the respondent asking him or her again to complete and return the survey questionnaire.

In other situations, respondents might be unavailable for several days or longer (not just at the time of contact by an interviewer or at the time a survey is delivered). Respondents may be on vacation or away on a business trip. As long as these temporary periods of unavailability do not extend beyond the field period of a survey, most survey firms hold these cases for an appropriate amount of time and then make additional attempts to contact these respondents and to obtain their cooperation.

There also are instances when the respondent is unavailable during the entire field period of a survey. These instances can include, for example, extended vacations and business travel to other countries. Unless the field period of the survey is extended, these cases usually are coded as permanently unavailable and are not processed further or contacted again.

Matthew Courser

See also Callbacks; Final Dispositions; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Unbalanced Question

An unbalanced question is one that has a question stem that does not present the respondent with all reasonably plausible sides of an issue. The issue of balance in a survey question also can apply to the response alternatives that are presented to respondents. Unbalanced questions generally are closed-ended questions, but it is possible to use open-ended questions in which the question stem is unbalanced. An unbalanced question will not always lead to biased data, but that is the concern in most instances.

For example, the following closed-ended question is unbalanced for several reasons and will lead to invalid (biased) data:

> *Many people believe that American troops should be withdrawn from Iraq as soon as possible. Do you Strongly Agree, Agree, Somewhat Agree, or Strongly Disagree?*

First, the question stem is unbalanced because it presents only one side of the issue in noting only one position taken by some in the general public. Second, the response alternatives are not balanced (symmetrical) as there are three "agree" choices and only one "disagree" choice. Third, the three response alternatives have no true midpoint; this is another aspect of the asymmetrical (unbalanced) nature of the response alternatives.

In contrast, a balanced version of this question would be as follows:

> *Some people believe that American troops should be withdrawn from Iraq as soon as possible, whereas other people believe that they should remain in Iraq until the country is more stable. What is your opinion on whether the troops should be withdrawn as soon as possible? Do you Strongly Agree, Somewhat Agree, Somewhat Disagree, or Strongly Disagree?*

This wording is balanced because it poses both sides of the issue. It also has a symmetrical set of response alternatives with two choices for "agree" and two similarly worded choices for "disagree." Furthermore, it has a true midpoint even though that midpoint does not have an explicit response alternative associated with it. If the researchers wanted to add a fifth choice representing the midpoint they could have added, "Neither Agree nor Disagree" in the middle.

In writing survey questions, it is easy for a researcher to avoid using unbalanced questions, unless there is a specific purpose to use such a question. A legitimate use would be the methodological testing of such question wording to study how such wording affects the answers given by respondents. In some other cases, an unethical or unscrupulous researcher may purposely use unbalanced questions to bias data in the directions that favor his or her client's interests. For example, a client with a particular political agenda can have a researcher word questions in an unbalanced fashion to make it more likely the answers respondents give are in a direction favored by the client. The unbalanced question—*Many people believe that American troops should be withdrawn from Iraq as soon as possible. Do you Strongly Agree, Agree, Somewhat Agree, or Strongly Disagree?*—makes it more likely that respondents will agree than if they were presented with a balanced version of this question. If the client wanted data to show that a larger proportion of the public wants troops to be withdrawn, then using an unbalanced question such as this would accomplish that end. Because too few reporters, editors, and producers in the news media ask to see the exact question wording that underlies survey data, they are not in a position to know whether the data are likely to be biased. As a result, data from unbalanced questions are easily disseminated without the news organization or the public knowing better.

Paul J. Lavrakas

See also Balanced Question; Closed-Ended Question; Open-Ended Question; Question Stem; Random Assignment; Response Alternatives

Further Readings

American Association for Public Opinion Research. (2007). *Question wording*. Retrieved March 6, 2008, from http://www.aapor.org/questionwording

Shaeffer, E. M., Krosnick, J. A., Langer, G. E., & Merkle, D. M. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly*, *69*(3), 417–428.

Unbiased Statistic

An unbiased statistic is a sample estimate of a population parameter whose sampling distribution has a mean that is equal to the parameter being estimated. Some traditional statistics are unbiased estimates of their corresponding parameters, and some are not. The simplest case of an unbiased statistic is the sample mean. Under the usual assumptions of population normality and simple random sampling, the sample mean is itself normally distributed with a mean equal to the population mean (and with a standard deviation equal to the population standard deviation divided by the square root of the sample size). A sample proportion is also an unbiased estimate of a population proportion. That is not surprising, as a proportion is a special kind of mean where all of the observations are 0s or 1s.

The matter is more complicated with regard to the sample variance. If the sum of the squared differences of the sample observations from the sample mean is divided by the sample size n, that statistic is *not* unbiased for estimating the population variance. To get an unbiased estimate of the population variance, the researcher needs to divide that sum of squared deviations by one less than the sample size.

The situation is even more complicated for the sample standard deviation. Although the sample variance obtained by dividing the sum of squares by $n - 1$ provides an unbiased estimate of the population variance, the square root of that statistic is *not* an unbiased estimator of the square root of the population variance (i.e., the population standard deviation), despite some claims made in certain statistics textbooks. (In mathematical statistical jargon, the expected value [mean] of the square root of a statistic is not, in general, equal to the square root of the expected value of the original statistic.)

For bivariate normal distributions for which the Pearson product–moment correlation coefficient (r) is a measure of the direction and the degree of linear relationship between the two variables, the sample r does not have a normal sampling distribution and is not an unbiased estimate of its population counterpart. The principal reason for this is that r is "boxed in" between -1 and $+1$. Fisher's z-transformation of r can be employed to partially remove the bias, and it is used frequently in testing hypotheses about population correlations and in establishing confidence intervals around sample correlations. The r is transformed to z (Fisher's z, not standardized variable z), and the test is carried out in that metric. For the confidence interval, r is transformed to z, the interval is obtained for z, and the end points of the interval are transformed back to the r scale.

In the survey sampling context, a well-known statistic for estimating the population total or mean in the presence of additional auxiliary information is the ratio estimator. Although this estimator is not unbiased for the corresponding population parameter, its bias is in part a decreasing function of the sample size.

It perhaps should go without saying, but an unbiased statistic computed from a given sample is not always equal to the corresponding parameter. Unbiasedness is strictly a "long run" concept. Thus, although not necessarily equal to the population parameter for any given sample, the expected value of the statistic across repeated samples is, in fact, the parameter itself. On any given occasion, a biased statistic might actually be closer to the parameter than would an unbiased statistic. When estimating a population variance, for example, division of the sum of squares by n rather than $n-1$ might provide a "better" estimate. That statistic is the "maximum likelihood" estimator of a population variance. So care should be taken when choosing a statistic based on its bias and precision; that is, variability.

Thomas R. Knapp

See also Bias; Design-Based Estimation; Model-Based Estimation; Population Parameter; Precision; Variance

Further Readings

Knapp, T. R. (1971). N vs. N-1. *American Educational Research Journal*, *7*, 625–626.

Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Undecided Voters

Undecideds, in election campaign lingo, are voters who have yet to decide which candidate—or side of an issue or referendum—they will support on Election Day. In pre-election polls, there are several ways of measuring support for a candidate or issue and, concomitantly, the number of undecided voters. A few researchers argue for an open-ended question format, in which the respondent volunteers the name of his or her preferred candidate. Answering such open-ended questions requires respondents to recall candidate names on their own, which means that these questions generally have larger percentages of "don't know" responses than closed-ended questions in which respondents hear or read candidates' names.

More frequently, however, researchers use a closed-ended "trial heat" question because the closed-ended measure has proven over time to be most accurate if comparison to Election Day results is the yardstick. These "read-list" questions in a telephone survey typically list all the candidates—or at least the major party candidates in a race—and ask whom the respondent would vote for "*if the election were held today*." Some research suggests that, especially in state and local races, it is important to list all candidates and their parties, rather than asking only about major-party candidates or about major-party candidates and the nonspecific ideal candidate, "someone else." For those who were initially undecided or refused to express a candidate preference, the question would attempt to ascertain whom the respondent leaned toward supporting. By combining these "leaners" with the original "choosers," the question determines overall support.

At least three other types of respondents are interviewed in pre-election polls: (1) those who say they will vote in the election but won't cast a vote in one race or another, (2) those who refuse to say whom they will support, and (3) those who are truly undecided. Often election pollsters combine the three types for reporting and analysis purposes into a category called "no opinion."

Proportions of Undecided Voters

Generally, the percentage of undecided voters in pre-election polls dwindles the closer to the election that the polls are taken. This certainly is the case in presidential elections. In local elections, however, the proportion of undecided voters often is quite large right up until Election Day. Any number of factors could explain the relatively large or changing size of undecided voters: low candidate name recognition, low salience of the election or issue, or events that cause a dramatic shift in support late in the campaign.

One such event is the death of the incumbent candidate during the campaign, especially when the death occurs late in the campaign, such as the 2002 plane crash that killed U.S. Senator Paul Wellstone (D-Minn.). Another such event would be the withdrawal of a candidate from the race because of ethical revelations late in the campaign. Some researchers also suggest that a large proportion of undecided voters late in the campaign may indicate a faulty likely voter model.

Examining demographic, geographic, political, and attitudinal profiles of undecided voters can provide useful information to campaign and party pollsters, who want to know what it might take to persuade undecided voters to support their candidate. Pollsters for news organizations do similar analyses because understanding undecided voters is an important part of the news story, particularly close to Election Day.

These analyses often reveal that undecided voters have lower socioeconomic status and are less connected to the political process than are decided voters. Politically, they are less likely than decided voters to be partisans, and they are often more likely to be moderates than liberals or conservatives. Demographically, they tend to have less education and lower incomes than decided voters.

Allocating Undecided Voters

Part of the analysis of undecided voters is the examination of several "what if" scenarios that ask the question, "How will undecided voters split on Election Day?" There are several methods of allocating undecided voters to candidates or to one side or the other of ballot issues such as constitutional amendments or referenda.

Proportional allocation of undecided voters simply means allocating them in the same proportion as overall candidate support. For example, suppose that Candidate A has 50% of the support in a pre-election poll and Candidate B has 40%, with 10% undecided. Dividing 50% by 90% (the total "decided" pool) gives Candidate A 56% of the support; dividing 40% by 90% gives Candidate B the remaining 4%, thus 44% overall.

Equal allocation means dividing the undecided vote evenly between or among the candidates. In the previous example, 5% of the undecided voters would be allocated to each of the two candidates, with Candidate A predicted to receive 55% of the vote and Candidate B 45%. Many researchers would be cautious about using this method in races where there are two major-party candidates and many minor-party candidates, simply because large numbers of undecided voters rarely break toward minor-party candidates.

More sophisticated *statistical modeling allocation*, such as discriminant analysis, can also be used. Discriminant analysis and other multivariate techniques first look at candidates' supporter profiles using researcher-specified variables such as party identification or gender. Then the technique allocates undecided voters to candidates based on how similar they are to the candidates' supporters on these variables. The technique usually involves using a random half of the sample to generate the statistical formula or model that will classify undecided respondents and using the other half to test the accuracy of the formula.

A fourth method of allocation, which some campaign researchers favor, has two steps. First, because party identification is so closely correlated with candidate choice, undecided partisan voters are allocated to their party's candidate—undecided Democrats would be allocated to the Democratic candidate and undecided Republicans, to the Republican candidate. Next, the remaining undecided voters without a party affiliation are allocated proportionally.

There are other less-empirical ways of allocating undecided voters. One is called the "incumbent rule," as advocated by pollster Nick Panagakis. Although popular references call it a "rule," it is less a formula than it is a conventional wisdom guideline adopted by some pollsters over the past two decades. The rule, which has received very little attention from political scientists, suggests that in an election where there is a well-known incumbent with less than 50% of the support and a well-known challenger, the majority of the undecided voters ultimately will vote for the challenger and against the incumbent. The rule assumes that the election is a referendum on the incumbent, and if voters do not back the incumbent after a term in office, they probably are not supporters. The rule appears to hold better in state and local elections than in presidential elections. Recently there has been some evidence that the incumbent rule may be weakening.

Robert P. Daves

See also Leaning Voters; Likely Voter; Pre-Election Polls; Trial Heat Question

Further Readings

Blumenthal, M. (2006, August 14). *Incumbent rule redux.* Retrieved December 26, 2006, from http://www.pollster.com/incumbent_rule/incumbent_rule_redux_1.php

Daves, R. P., & Warden, S. (1995). Methods of allocating undecided respondents to candidate choices in pre-election polls. In P. J. Lavrakas, M. W. Traugott,

P. V. Miller (Eds.), *Presidential polls and the news media* (pp. 101–122). Boulder, CO: Westview Press.
Panagakis, N. (1989, February 27). Incumbent races: Closer than they appear. *The Polling Report*, *5*, 1–3.
Vanishing Voter Project. (2006). *One in seven likely voters still undecided on a presidential candidate.* Retrieved March 8, 2008, from http://www.ksg.harvard.edu/presspol/vanishvoter/VV2000/Press_Releases/10-18-00.shtml

UNDERCOVERAGE

Undercoverage occurs when an element of the target population is not represented on the survey frame and therefore not given any chance of selection in the survey sample; that is, the element has zero probability of selection into the sample. Undercoverage is the most serious type of coverage error because it can be difficult to detect and even more difficult to solve. Therefore, preventing undercoverage is often a priority during survey design. Large survey operations often plan and budget for extensive coverage evaluations. For example, a large sample survey called the Accuracy and Coverage Evaluation was conducted by the U.S. Census Bureau during Census 2000, with separate staff and separate physical office space. Its primary purpose was to evaluate the coverage of the census.

In household surveys of the general population, there are generally two levels at which undercoverage is a concern. First, households may be missing from the frame. For example, in a random-digit dialing telephone survey, households without telephone service will be missed. In addition, while landline telephone coverage of most U.S. populations had been increasing, the recent popularity of cell phones has begun to jeopardize coverage of traditional landline telephone frames because of the rapidly growing cell phone only population—more than 20% of all U.S. households as of 2008. In surveys that use an area frame, households can be missed during the listing operation for a variety of reasons; for example, difficult to visually spot, misclassified as vacant or business, incorrectly assigned to across a boundary into an unsampled geographic area.

Second, even when a household is on the frame, some people within the household may not be covered. Unfortunately, this type of undercoverage cannot be prevented by good frame construction. There are two major theories on why people fail to identify with the household. Roger Tourangeau found that the initial contact person at the household underreported the number of household members when he or she was asked for a list with several pieces of information on each person. In contrast, when asked for only each person's initials, more household members were reported. The authors theorized that this was because of privacy concerns. Elizabeth Martin found that in many cases the initial contact person at the household lacked enough knowledge about other household members to accurately include or exclude them according to the survey's residency rules. Both studies found issues tended to arise under complicated living arrangements (multiple roommates, group housing, recent changes, etc.).

There are several ways to prevent undercoverage. The survey should use a frame that provides the necessary unit coverage. That is, a survey that is estimating mobile phone use should not use a traditional landline telephone frame. For surveys that use lists for frame construction, the lists should be as recent as possible. In business surveys, lists of establishments or enterprises can quickly become out of date due to mergers, acquisitions, and new openings or closings. For surveys that use an area frame, quality checks can be performed on the address listing operation (i.e., "re-listing") using a subsample of areas or more expert staff.

There are both simple checks for undercoverage and extensive studies of undercoverage. Simple checks include comparing simple survey estimates (e.g., demographics) to other sources, such as recent prior surveys or census data. Using external data, post-survey adjustments to the analysis weights can be made if necessary.

Some surveys budget for extensive undercoverage evaluations, but this is not practical for all surveys. Dual-frame estimation can be used if there is an independent data collection and survey elements can be matched. Other data outside the survey can be analyzed for verification. For example, J. Gregory Robinson evaluated the coverage of the 1990 U.S. census using birth, death, immigration, and emigration records. This type of analysis can give useful coverage estimates in different demographic categories.

Jeffrey M. Pearson

See also Address-Based Sampling; Area Frame; Cell Phone Only Household; Coverage Error; Elements; Frame; Target Population; Unit Coverage; Within-Unit Coverage

Further Readings

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys.* New York: Wiley.

Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, *63*(2), 220–236.

Pearson, J. M. (2005). Quality and coverage of listings for area sampling. *2005 Proceedings of the American Statistical Association, Section on Government Statistics* [CD-ROM]. Alexandria, VA: American Statistical Association.

Robinson, J. G., Ahmed, B., Gupta, P. D., & Woodrow, K. A. (1993). Estimation of population coverage in the 1990 United States Census based on demographic analysis. *Journal of the American Statistical Association*, *88*(423), 1061–1071.

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? Survey undercoverage and household roster questions. *Journal of Official Statistics*, *13*(1), 1–18.

U.S. Census Bureau. (2004). *Accuracy and coverage evaluation of Census 2000: Design and methodology*. Washington, DC: Author.

UNDERREPORTING

When answering questions on sensitive behaviors, many respondents show a specific variant of response error: They tend to report fewer instances of undesired behaviors compared to what they have actually experienced. This is called *underreporting*. It is assumed that respondents avoid reporting unfavorable conduct because they do not want to admit socially undesirable behaviors, which in turn leads to this type of misreporting. Similar effects are known for responses to survey questions about unpopular attitudes.

Currently, it is not known whether underreporting is the result of a deliberate manipulation of the true answer or whether it occurs subconsciously. Nevertheless, for the most part it is assumed to be a response error that is associated to the cognitive editing stage of the question–answer process. Misreporting due to forgetting or other memory restrictions is considered to be a minor source of underreporting.

Several validation studies use records of the true answers to provide evidence for underreporting. For example, studies on abortion or illegal drug use show that fewer respondents admit those behaviors, or if they admit them, they report fewer instances of the presumed undesired behavior.

Since underreporting is related to social desirability bias it occurs more often in interviewer administered settings compared to self-administered surveys. Accordingly, audio, audiovisual, and telephone audio computer-assisted self-interviewing methods have been used extensively when collecting responses on socially undesirable behaviors or attitudes. These modes reduce the response burden of answering sensitive questions, which in turn is assumed to reduce underreporting of undesirable behaviors or attitudes.

Also, several question techniques have been developed to reduce underreporting. Besides forgiving question wording, the "everybody approach," and indirect questioning, the sealed ballot method has proven to effectively reduce underreporting. In addition, several variants of the randomized response technique are available. However, because the instructions necessary for those question types are rather complicated and not always understood by the respondent, they are practically seen as less effective in reducing underreporting.

Marek Fuchs

See also Audio Computer-Assisted Self-Interviewing (ACASI); Overreporting; Randomized Response; Response Error; Self-Administered Questionnaire; Sensitive Topics; Social Desirability

Further Readings

Magura, S., & Kang , S.-Y. (1996). Validity of self-reported drug use in high risk populations: A meta-analytical review. *Substance Use and Misuse*, *31*(9), 1131–1153.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

UNFOLDING QUESTION

An unfolding question refers to a type of question sequence that yields more complete and accurate data than would a single question on the same topic. Unfolding questions are used by survey researchers in an attempt to reduce item nonresponse (i.e., missing data) and measurement error.

For example, asking someone into which of the following income categories her 2007 total household income fell—less than $20,000; $20,000–$39,999; $40,000–$59,999; $60,000–$79,999; $80,000–$99,999; $100,000 or more—will lead to a good deal of "Don't Know" or "Refused" answers. Researchers have found that an unfolding question about income will substantially reduce item nonresponse and thus the need to impute those missing values.

An unfolding income question sequence for the income variable referenced in the preceding paragraph, that was programmed to be asked in a computer-assisted interview, would be as follows:

Q1. Was your total household income from all sources in 2007 more or less than $60,000?
< 1 > MORE (GO TO Q4)
< 2 > LESS (GO TO Q2)
< 3 > $60,000 (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)

Q2. And was it more or less than $40,000?
< 1 > MORE (GO TO Q6)
< 2 > LESS (GO TO Q3)
< 3 > $40,000 (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)

Q3. And was it more or less than $20,000?
< 1 > MORE (GO TO Q6)
< 2 > LESS (GO TO Q6)
< 3 > $20,000 (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)

Q4. And was it more or less than $80,000?
< 1 > MORE (GO TO Q5)
< 2 > LESS (GO TO Q6)
< 3 > $80,000 (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)

Q5. And was it more or less than $100,000?
< 1 > MORE (GO TO Q6)
< 2 > LESS (GO TO Q6)
< 3 > $100,000 (GO TO Q6)
< 8 > REFUSED (GO TO Q6)
< 9 > UNCERTAIN (GO TO Q6)

In this example, after the income sequence has been administered, all respondents are taken to Q6 (the next logical question topic). Of note, even though the entire unfolding sequence comprises five questions, any one respondent would only be asked one, two, or three of the questions, not all five of them and the majority would only be asked one or two. This five-question unfolding sequence will lead to far fewer missing income values than a single income question that presents essentially the same income categories all at once. It also will yield data that the researchers can combine to form a single income variable with the desired six categories in the original one-question income example.

Unfolding questions can be asked for other topics that are measured on some form of ordered scale. They are particularly useful when there are many response choices on the scale and when the cognitive burden on many respondents is too great to present all the choices at once. If all choices were presented at once, primacy, recency, and other satisficing effects would likely lead to errors in the data. For example, a scale measuring political orientation that uses the eight response options (e.g., Extremely Conservative, Very Conservative, Conservative, Somewhat Conservative, Neither Conservative Nor Liberal, Somewhat Liberal, Liberal, Very Liberal, Extremely Liberal) would gather more accurate data if it were presented as an unfolding question sequence starting with a question asking, "*In terms of politics, are you liberal or conservative?*"

Paul J. Lavrakas

See also Imputation; Measurement Error; Missing Data; Primacy Effect; Recency Effect; Respondent Burden; Satisficing

Further Readings

Duncan, G. J., & Petersen, E. (2001). The long and short of asking questions about income, wealth, and labor supply. *Social Science Research*, *30*(2), 248–263.

UNIT

Sampling, variance, analysis, reporting, and dissemination units cover most of the unit types of interest in survey methodology. *Sampling units* can be primary, secondary, tertiary, and beyond—also known as first-stage, second-stage, and so on. These refer to a hierarchy of smaller and smaller geographic or organizational structures that have been exploited in the sample design to improve efficiency. In face-to-face surveys of the general population, the stages are typically county (or count group), block (or cluster of neighboring blocks), housing unit (or other quarters), and person. In student surveys, the stages are typically school district, school, classroom, and student. In random-digit dialing surveys, the stages are typically just household and person, although sometimes a 100-bank is also a stage.

Variance units are often the same as primary sampling units (PSUs) but can be different. Sometimes PSUs are collapsed into super PSUs to reduce the number of replicate weights that need to be created for variance estimation via resampling methods such as the bootstrap or jackknife. Other times, PSUs are split into pseudo PSUs to improve the stability of variance estimates even if it means a downward bias in the variance estimates. Both operations can also be done to make it harder for data snoopers to discover anything about known sample members from public use files.

Analysis units are often people but can be households or families. However, these other constructs usually are well-defined only at a particular point in time. Efforts to define stable interpersonal groupings over time have been largely unsuccessful, in the United States at least. Even the concept of "parent" is hard to keep stable over time, as varying constellations of adults can provide parenting services to children during their minor years. Where family structures and parents have been of interest over time, the most successful approach has been to follow the persons of interest and report the characteristics of their families or parents as attributes of themselves.

Reporting units are generally the same as ultimate stage sample units, but in surveys of businesses and other larger social constructs such as governmental bodies, a "large" sample unit may need to be split into multiple reporting units. For example, in a survey of energy usage in commercial buildings, the individual tenants might be the reporting units for a multitenant building with separately metered utilities.

Dissemination units are the smallest geographic or organizational structures for which formal estimates are published. Typical examples in the United States include the four census regions, the states, and counties, depending on the size of the survey. Sometimes, there is interest in the primary sampling units as analysis or dissemination units. Metropolitan areas constitute one example where attempts have been made to use them both as sampling units and as dissemination units. These attempts are not, however, very satisfying. Sample sizes are generally too small within individual PSUs to satisfy user expectations for the precision of formally published estimates. Also, it is hard to explain why, if such estimates are important or useful, they are only available for a few dozen of them.

David Ross Judkins

See also Level of Analysis; Level of Measurement; Primary Sampling Unit (PSU)

Further Readings

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Asok, C. (1984). *Sampling theory of surveys with applications* (3rd ed.). Ames: Iowa State University Press.

Westat. (2000). *WesVar 4.0 user's guide* (App. D). Rockville, MD: Author.

UNIT COVERAGE

The proportion of the intended referential universe that is represented (on a weighted basis) in the sampling frame of a survey is referred to as the *coverage rate*. Unit coverage refers to how extensively the units (e.g., households, schools, businesses) in the universe are "covered" by (i.e., included in) the sampling frame, which includes the dynamic process whereby an enumerator (lister) or interviewer decides whether a household or person should be included in the frame.

For purposes of setting public policy, it is generally considered desirable to hear the voices of a broad swath of the public. Even in voter intention surveys, pollsters are more confident in their predictions if their screening survey for likely voters covers most of society. As such, survey researchers should strive to use sampling frames that cover the units of the population as well as is possible.

Coverage is a function of both how a sample is selected and how contact with sample units is attempted. Being "covered" does not necessarily mean that the person responds. But it does include whether they "cooperate" enough to confirm their existence. For example, an interviewer may approach a house that appears to be abandoned and knock on the door. If the occupant shouts out a demand that the interviewer leave, then he or she is covered. If, on the other hand, the occupant remains stealthily silent so that the interviewer erroneously classifies the house as vacant, then he or she is not covered. As another example, a person who answers the phone when contacted as part of a random-digit dialing survey and then hangs up without saying anything else is covered, but a person who never answers the phone because of a reaction to the caller ID is typically not covered. Errors in coverage that lead to households or people being "missing" from the frame make up *undercoverage.*

There are a variety of reasons for undercoverage. One is that some members of society are secretive, antisocial, or both. Examples include illegal migrants, squatters, fugitives, members of insular religious cults, and members of rebellion-like movements. Another reason is mental or physical illness that makes it impossible for someone to communicate with a survey interviewer. A third is high mobility, as in the case of migrant farm workers. A fourth is physical remoteness of dwellings. A fifth is shared addresses and improvised apartments. Electronic connectedness matters for some surveys. The desire to protect children also can be a factor. For example, it is known that overtly screening a sample for households with children can lead to much lower coverage of children than a general interview where information about children is collected as a secondary issue.

A continuum of eagerness and capacity for social engagement can range from schizophrenic homeless wanderers and misanthropic hermits at one extreme to politicians who live almost purely in the public eye at the other. Professional experience in the United States indicates that people in the top half of that continuum can be covered easily, provided that the survey has a legitimate social research purpose. People in the second quartile of the continuum require more effort but can still usually be covered with fairly low-cost methods. People between the 15th and 25th percentiles of the continuum can usually only be covered with fresh area samples conducted by door-to-door interviewers under government auspices. People between the 5th and 15th percentiles usually can be covered only in official surveys conducted by the Census Bureau. People between the 2nd and 5th percentiles can be covered only through heroic means, such as those employed in the decennial censuses. People in the first two percentiles are mostly not coverable by any procedure. Or rather, the extraordinary procedures that are required to try to cover them also result in *overcoverage* where some other people are counted multiple times, so that it becomes practically impossible to determine the coverage rate.

Estimating unit coverage is difficult, but there are some fairly well-accepted conventions. In the United States, coverage measurement for demographic surveys usually starts with official population estimates published by the Census Bureau. These are synthesized from the count from the most recent decennial census, vital statistics, emigration and immigration statistics, tax statistics, and school enrollment statistics. A historical examination of coverage rates requires careful attention to benchmark methodology, as there have been occasional major shifts. One example concerned the counting of illegal aliens in the 1980s. Another concerned a period in the 1990s when there were parallel series: one adjusted for undercount in the decennial census, the other not.

Another difficulty in estimating coverage rates comes in separating undercoverage from nonresponse. If a combined adjustment is made for nonresponse and undercoverage, more effort is required to determine the coverage rate. A procedure preferred by some organizations is to use separate nonresponse and undercoverage adjustments. Weighted totals of the population before and after the undercoverage adjustment can then be compared to estimate the coverage rate.

In designing a survey, consideration must be given to the likely association between variables of interest and social attachment. Generally speaking, little is known about uncovered persons, so decisions about the targeted coverage rate must be based mostly on the survey designer's judgment. Surveys of antisocial behavior require the best coverage to be credible.

David Ross Judkins

See also Coverage; Nonresponse; Overcoverage; Undercoverage

Further Readings

Chromy, J. R., Bowman, K. R., Crump, C. J., Packer, L. E., & Penne, M. A. (1999). Population coverage in the National Household Survey on Drug Abuse. *Proceedings of the Section on Survey Research Methods* (pp. 575–580). Alexandria, VA: American Statistical Association.

Horrigan, M., Moore, W., Pedlow, S., & Wolter, K. (1999). Undercoverage in a large national screening survey for youths. *Proceedings of the Section on Survey Research Methods* (pp. 570–575). Alexandria, VA: American Statistical Association.

Judkins, D., DiGaetano, R., Chu, A., & Shapiro, G. (1999). Coverage in screening surveys at Westat. *Proceedings of the Section on Survey Research Methods* (pp. 581–586). Alexandria, VA: American Statistical Association.

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? Survey undercoverage and household roster questions. *Journal of Official Statistics*, *13*(1), 1–18.

UNIT NONRESPONSE

Unit nonresponse in a survey occurs when an eligible sample member fails to respond at all or does not provide enough information for the response to be deemed usable (not even as a "partial completion"). Unit nonresponse can be contrasted to item nonresponse (missing data) wherein the sample member responds but does not provide a usable response to a particular item or items. Unit nonresponse can be a source of bias in survey estimates, and reducing unit nonresponse is an important objective of good survey practice.

Reasons for Unit Nonresponse

Despite the best efforts of the survey practitioner, there are reasons why unit nonresponse will still occur.

Refusal: The sample member may refuse to participate in the survey. Often refusals are divided into *hard* refusals and *soft* refusals, depending on the intensity with which the sample member refuses to participate.

Sample member cannot be found: The sample member (whether a person, a household, or an establishment) may have moved or otherwise cannot be located. In the case of a telephone survey, the telephone may ring repeatedly with no answer or be picked up by an answering machine with an uninformative message.

Sample member may be temporarily away: The sample member may be known to be travelling and unavailable during the data collection period.

Communication difficulty: The sample member may speak a language that none of the interviewers speaks or into which the survey instrument has not been translated. In other instances, the sample member may have a physical or mental disability that interferes with communication.

Sample member provides inadequate data: The sample member provides some data, but after editing, the data are deemed inadequate to constitute a valid response.

Other: There are other reasons for unit nonresponse, such as quarantines, which although unlikely, do occur. (Detailed discussion of these other instances is available at www.aapor.org.)

It is important to distinguish sample members that are unit nonrespondents from those that are ineligible (out of scope). In the case of sample members that cannot be contacted, it may not be possible to know for certain if a particular sample member is a nonrespondent or ineligible, but the proportion of nonrespondents among the "noncontacts" can be estimated. An example is a random-digit dial telephone survey of households in which a selected telephone number is called repeatedly with no answer. The telephone number may belong to a household or an establishment, or it may not be an assigned number. Only in the first case is the sample member a unit nonrespondent.

Reducing Unit Nonresponse

Steps that may be taken to reduce unit nonresponse include the following:

- Keep the survey instrument (questionnaire, Web form, or interviewer protocol) short and relatively easy to complete. It helps if the items are interesting and salient.
- Use only thoroughly trained interviewers.
- Provide an advance letter or telephone call or other message. This helps the sample member to distinguish serious surveys from marketing and promotions. Including endorsements from respected organizations can also be beneficial.
- Provide a cash or other type of incentive.
- Allow proxy respondents. If the intended sample member is unavailable or uncooperative, someone else (e.g., a family member) may be able to respond for the person.
- Provide the sample member a choice of modes with which to respond. For example, the sample member may prefer to respond on the Web or by telephone rather than by mail in a paper questionnaire. Telephone follow-up to mail questionnaires is a frequently employed technique, to allow the respondent to be interviewed at that time.
- Assure the sample member of the confidentiality of the responses.

- If the sample member initially refuses to respond, make use of advanced refusal conversion techniques.
- Employ intense and well-designed follow-up techniques. For example, in telephone surveys, a large number of callbacks may be needed at varied times of the day and days of the week.
- In longitudinal surveys, make use of sophisticated tracking techniques to track sample members who may have moved between rounds of the survey.

Efforts to reduce unit nonresponse can have drawbacks. Incentives and follow-up activities are costly. If the follow-up is too intense, survey members may provide low-quality responses so as not to be bothered further. Proxy respondents may not be able to provide data as accurate as the intended respondent. In some cases, the disadvantages may even override the benefits of the reduced unit nonresponse, in part because not all unit nonresponse leads to unit nonresponse error.

In many surveys there are "hard-core" nonrespondents that are almost impossible to get to respond. In some cases, unfortunately, these nonrespondents may be of special interest. In household travel surveys, for example, there is great interest in recent immigrants because these individuals may not have obtained a driver's licence and hence may be dependent on public transportation, walking, and bicycling as modes of transportation. Yet because of possible language barriers, lack of telephone coverage, and other factors, recent immigrants may be especially difficult to get to respond.

Adjusting for Unit Nonresponse

Post-survey adjustments are usually made to ameliorate the effects of unit nonresponse. It is typically the case that unit nonresponse is not distributed evenly in the sample. In surveys of persons, for example, the rate of unit nonresponse is usually greater for younger adults than for older adults, for minorities than for whites, and especially among those with lower educational attainment compared to those with higher educational attainment. In establishment surveys, larger establishments may be more likely to respond than smaller ones. The most commonly employed adjustment is the weighting class method for adjusting the sample weights. The sample members are divided into groups, called *weighting classes*, which are deemed to be relatively homogeneous. Within a weighting class, the weights of all the unit respondents are multiplied by the same factor (at least one) so that they sum to the estimated population size for the weighting class.

The propensity scoring method of adjustment uses information available for all sample members to estimate each sample member's probability of responding. These probabilities of responding ("propensities") are then used to adjust the sample weights of the responding units.

Another adjustment method sometimes used is to impute all the items (or all the key items) for a unit nonrespondent. This method is particularly common in census data collections (where all sample members are selected with certainty), where sample weights are not employed.

Sometimes unit nonrespondents (at a particular point in time) are subsampled for intensive follow-up. The weights of the units in the subsample that eventually respond are increased to account for the subsampling. This procedure controls nonresponse bias at the expense of an increase in variance.

Michael P. Cohen

See also Advance Contact; Confidentiality; Eligibility; Incentives; Language Barrier; Missing Data; Noncontacts; Nonresponse; Nonresponse Bias; Partial Completion; Propensity Scores; Proxy Respondent; Refusal; Refusal Conversion; Respondent Burden; Unable to Participate; Unit; Weighting

Further Readings

Groves, R. M. (2004). *Survey errors and survey costs.* New York: Wiley.

Kalton, G. (1983). *Compensating for missing survey data.* Ann Arbor: University of Michigan, Institute for Social Research.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys.* New York: Wiley.

Unit of Observation

A unit of observation is an object about which information is collected. Researchers base conclusions on information that is collected and analyzed, so using defined units of observation in a survey or other study helps to clarify the reasonable conclusions that can be drawn from the information collected.

An example of a unit of observation is an individual person. Other examples include a family or a neighborhood.

A survey or other type of study can involve many different levels of units of observation. For example, the U.S. Census 2000 used a hierarchical arrangement to describe the units of observation about which it collected information. These units range from "United States" to "region" to "census block."

Some researchers distinguish between the terms *unit of observation* and *unit of analysis*. For example, a unit of observation might be an individual person, but a unit of analysis might relate to the neighborhood in which the individual lives, based on data collected about individuals in the neighborhood. Clear articulations of units of observation through use of specific definitions lend clarity to survey and study efforts.

Focusing on the unit of observation throughout the course of a study—from inception to conclusions to dissemination of results—helps researchers later present an organized explanation of the phenomenon and helps keep the explanation relevant to the data collected.

Heather H. Boyd

See also Case; Unit

Further Readings

U.S. Census Bureau. (2003, June). *2000 Census of population and housing: Summary of social, economic, and housing characteristics, selected appendices, PHC-2-A*. Retrieved February 9, 2007, from http://www.census.gov/prod/cen2000/phc-2-a.pdf

Universe

The universe consists of all survey elements that qualify for inclusion in the research study. The precise definition of the universe for a particular study is set by the research question, which specifies who or what is of interest. The universe may be individuals, groups of people, organizations, or even objects. For example, research about voting in an upcoming election would have a universe comprising all voters. If the research was about political parties' sponsorship of candidates, the research would include all political parties. In research that involves taking a sample of things for testing or review, the universe may include inanimate objects. For example, a researcher may want to determine the extent to which toys made in a particular country contain a certain type of paint. In this case, the universe would include all toys made in that country. Some survey research has a very small universe (e.g., a survey of individuals who have won 1 million dollars or more in the lottery), and some research has a large universe (a survey of the television-viewing habits of all adults in the United States).

Survey practitioners often do not expect to be able to measure the entire universe, because it is neither practical nor necessary. A survey of all the elements in a universe would not be a sample but would instead be a census. Taking a census is rarely practical because it is costly and it rarely is needed; that is, a good survey can gather as accurate (and sometimes more accurate) data about the universe.

Defining the universe is a key element in the design of an appropriate sampling strategy. The researcher must carefully consider precisely who or what is of interest and then how best to contact or include these individuals or things in the research. The importance of appropriately defining the universe should not be underestimated. A universe that has been too narrowly defined will exclude important opinions and attitudes, and a universe too broadly defined will include extraneous information that may bias or affect the overall results.

Once defined, the universe is used by the researcher to structure an appropriate sampling methodology, in particular by defining the target population and choosing a sampling frame or frames. Research being conducted with a universe of all adults who watch television, for example, can contact a random sample of such individuals using random-digit dialing or by sampling clusters of geographies and then sampling households within those geographies using addressed-based sampling or area probability sampling. Sometimes multiple frames are needed to adequately cover the universe. In many surveys, the researcher does not know (nor does she or he need to know in advance) how many individuals or things are qualified to be part of the universe. One way to estimate this information is to retain information from any screener used to identify and select qualified respondents. For example, imagine a researcher wanted to do a survey of all individuals who own *both* a dog and a cat in Australia. It is unlikely that there is any reliable information on the precise number of Australians who own both types of pets, and it is further unlikely that a listing of such pet owners exists. When screening randomly selected households, the researcher can determine how many of the individuals contacted qualify and can use this information to estimate the total universe size. For example, if the researcher randomly contacts 1,000 households and finds that 45 of these own both a dog and a cat, the researcher can then estimate that roughly between 3% and 6% of all households in Australia own both types of pets.

The universe sometimes is known as the *population of interest* or *target population*.

Sarah Butler

See also Address-Based Sampling; Area Probability Sample; Census; Population; Population of Inference; Population of Interest; Random-Digit Dialing (RDD); Research Question; Sampling Frame; Screening; Target Population

Unknown Eligibility

Unknown eligible is a category of survey dispositions that is used in all survey modes (telephone, in-person, mail, and Internet) when there is not enough information known about the case to determine whether the case is eligible or ineligible. For this reason, most survey organizations make additional contact attempts on cases categorized into one of the unknown eligibility categories to try to definitively categorize the case into either the eligible or ineligible category. Because the unknown eligible category is broad, researchers usually use a more detailed set of survey disposition codes to better categorize the outcomes of cases that fall into the unknown eligible category.

In a telephone survey, cases are categorized as unknown eligibles under two circumstances: (1) when it is not known whether an eligible household exists at a telephone number and (2) when it is known that a household exists at the number, but it is unknown whether an eligible respondent exists at the number. For cases in which it is not known whether a household exists at the sampled telephone number, cases of unknown eligibility include those numbers that were included in the sample but never dialed by an interviewer. Another common subcategory of cases of unknown eligibility include numbers that always ring busy, numbers that reach ambiguous answering machine or voicemail messages that do not indicate whether a household has been reached, privacy manager technologies that block interviewers from completing dialing attempts, and telephone line–related technical problems, such as "all circuits are busy." If it is known that a household exists at the sampled telephone number, cases of unknown eligibility also can occur when it is not possible to complete a screener, and thus it is not known whether a household has an eligible respondent.

In an in-person survey, cases are categorized into the unknown eligibility category when one of two circumstances exists: (1) when it is not known whether an eligible household exists at an address and (2) when it is known that a household exists at the sampled address, but it is not known whether an eligible respondent lives at the address. For cases in which it is not known whether a household exists at the address, cases of unknown eligibility include those in which a household was not contacted by an interviewer (i.e., the case was not worked), cases in which an interviewer was unable to reach the household (most commonly due to causes related to safety and building security), and cases in which an interviewer was not able to locate the sampled address. If it is known that a household exists at the sampled address, cases of unknown eligibility also can occur when it is not possible to complete a screener, and thus it is not known whether a household has an eligible respondent.

Cases of unknown eligibility occur for five principal reasons in mail surveys of specifically named respondents. One reason that a case of unknown eligibility might occur in a mail survey is that the questionnaire was never mailed out, and thus, nothing is known about the sampled address. A second, more common, reason for a case in a mail survey being included in the unknown eligibility category occurs when the questionnaire was mailed out but the addressee did not receive the mailing. These cases include "refused by addressee," cases in which the U.S. Postal Service will not deliver mail to the addressee, and cases in which the address on the outer envelope is incomplete or illegible. A third type of cases of unknown eligibility occurs in mail surveys when a case reaches a sampled address, but it is not known whether an eligible respondent resides at the address; this most often occurs when a screener is not completed. A fourth type of cases of unknown eligibility occurs in mail surveys when a questionnaire is returned as undeliverable but includes forwarding information. The fifth, and most common, case is those mailings for which the researcher never receives back any information on whether or not the mailing was received, including no returned questionnaire.

Cases of unknown eligibility occur for two principal reasons in an Internet survey: (1) when an invitation to complete the survey is sent, but nothing is known about whether the invitation was received, and (2) when the researcher learns that the invitation was never delivered. In the former case, many invitations to complete an Internet survey are sent via email, and researchers may receive no notification if the email address the invitation was sent to is incorrect, not working, or is not checked by the sampled respondent. In the latter case, researchers may receive a message that indicates that the email address is incorrect and that the invitation cannot be delivered. In either instance, usually nothing is known about the sampled respondent and his or her eligibility to complete the survey.

There tends to be variance across survey organizations in how cases of unknown eligibility are treated in computing surveys response and cooperation rates. The most conservative approach is to treat these cases as eligible cases; the most liberal approach, which makes the survey appear more successful, would treat cases of unknown eligibility as ineligible and thus exclude them from the response rate calculations. In most instances, neither of these extremes is prudent, and instead researchers often treat a proportion of unknown eligibility cases (referred to as "*e*") as being eligible in their response rate calculation. When doing this, the researcher should have an empirical basis upon which to calculate the proportion counted as eligible or ineligible.

Matthew Courser

See also *e*; Final Dispositions; Ineligible; Out of Sample; Response Rates; Temporary Dispositions

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research.* Chicago: University of Chicago Press.

UNLISTED HOUSEHOLD

An unlisted household is a residential unit that does not have a telephone number listed in published directories or one that would be given out by directory assistance. A telephone number may be unlisted for several reasons. Household residents actively may have chosen to keep their telephone number unlisted. Alternatively, although the residents of an unlisted household may not have specifically chosen to keep their telephone number unlisted, the number may be too new to be found in published directories or via directory assistance. This can be a result of a recent move or change in telephone number.

Although the literature suggests important differences between the characteristics of respondents who choose to keep their telephone numbers unlisted and respondents whose telephone numbers are unlisted for other reasons, consistent findings have been reported on the larger differences between the characteristics of households with listed numbers versus those with unlisted numbers. Directory-listed sampling frames have tended to exhibit an overrepresentation of (a) established households, (b) middle- and higher-income households, (c) two or more person households, (d) older householders, (e) married householders, (f) better-educated householders, (g) retired householders, (h) white-race households, and (i) home owners.

In general, sampling frames based on listed telephone numbers exclude unlisted households. This coverage shortfall has the potential to introduce bias into surveys that use a directory-listed sampling frame to construct a representative sample of the general population of telephone households. The coverage issue is most acute in urban areas and the western region of the United States, where unlisted telephone rates are higher. Additionally, as the proportion of households without listed telephone numbers in the general population increases, the potential for bias due to their exclusion from directory-listed sampling frames is likely to increase commensurately. Random-digit dialing (RDD) is an alternative sample design that can overcome this coverage limitation. However, although RDD sampling frames include both listed and unlisted households, the higher proportion of nonworking numbers in RDD sampling frames can reduce the efficiency of the sample as compared to that of a directory-listed sampling frame. It also may reduce the external validity of the sample, as respondents with an unlisted number are more prone to survey nonresponse due to both noncontact and refusal.

Adam Safir

See also Coverage Error; Directory Sampling; External Validity; List-Assisted Sampling; Noncontacts; Random-Digit Dialing (RDD); Refusal; Sampling Frame; Unpublished Number

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author. Retrieved April 24, 2008, from http://www.aapor.org/uploads/standarddefs_4.pdf

Brick, J. M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, *59*, 218–235.

Currivan, D. (2003). Sampling frame. In M. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The SAGE encyclopedia of social science research methods.* Thousand Oaks, CA: Sage.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Tucker, C., Lepkowksi, J., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly*, *66*, 321–338.

UNMATCHED NUMBER

An unmatched telephone number is one that does not have a mailing address associated with it. Typically, it also does not have a name matched to it. The vast majority of unmatched telephone numbers are also unlisted telephone numbers. The proportion of unmatched numbers in a telephone survey that uses matching as a technique will depend upon the extent to which the survey organization uses multiple vendors to do their matching and the quality of the matching techniques and databases the vendor uses. However, in random-digit dialing surveys in the United States, regardless of how many matching vendors are used, a large minority of numbers can never be matched to an address because that information simply is not accessible to any matching company. This is due to the privacy and confidentiality concerns of the households whose numbers cannot be matched to an accurate address.

Whether a telephone number can be "matched" to an address is predictive of the likelihood that a completed interview will be attained with that household in a telephone survey. A greater proportion of interviews are completed with numbers that are matched than are completed with unmatched numbers. A primary reason for this is that those with unmatched numbers are less likely to react positively when they are contacted by a stranger (i.e., the interviewer). Another primary reason for this is that matched numbers have an address associated with them. As such, researchers can send advance mailings to these households when they are sampled for a telephone survey to alert them ("warm them up") to the fact that an interviewer will be calling them. This cannot be done with unmatched numbers. Instead, calls to them are "cold calls," and response rates for them consistently have been shown to be lower than with matched numbers.

On average, unmatched telephone numbers require more callbacks than do matched numbers in order for them to reach a proper final disposition. Thus, the calling rules used by a survey center to process unmatched numbers should differ from the rules used to process matched numbers. However, unless a survey center has their telephone samples screened for matched/unmatched status or receives this information for each number in their sample from their sample vendor, it will not be possible for them to take the matched/unmatched status into account as their computer-assisted telephone interviewing system processes the callback attempts.

Paul J. Lavrakas

See also Advance Contact; Advance Letter; Calling Rules; Cold Call; Computer-Assisted Telephone Interviewing (CATI); Listed Number; Matched Number; Random-Digit Dialing (RDD); Telephone Surveys; Unpublished Number

UNPUBLISHED NUMBER

The unpublished number disposition is used both in random-digit dialing (RDD) surveys of the general public and in telephone surveys of named people that use a list-based sample. Unpublished telephone numbers are numbers that are not listed in the local telephone directory, and the vast majority of these numbers also are unlisted and thus not available from directory assistance.

The unpublished number disposition is fairly rare in RDD surveys and is used only when an interviewer dials the telephone number of a case in the sampling pool and receives a recorded message from the local telephone company indicating that the telephone number dialed by the interviewer has been changed to an unpublished number—meaning that the original telephone number in the sampling pool is no longer in service. In RDD surveys of the general public, the new, unpublished number would not be called even in the unlikely possibility that it was known, because doing this would change the household's probability of selection to be included in the sampling pool—essentially doubling the household's chances of being included in the sample. Moreover, because RDD sampling techniques have a nonzero probability of reaching any household in a sampling area that has telephone service, the new, unpublished number also might be included in the sample already. In these cases, the unpublished number disposition would be considered a final disposition.

The unpublished number disposition is used more commonly in telephone surveys that use list-based sampling techniques. For example, if a researcher is sampling households from an address frame and then using reverse directory techniques to match telephone numbers to addresses in the sampling frame, there may be no telephone number available because the household or respondent may have designated his or her number as unpublished. In some cases the researcher may be able to contact directory assistance and obtain a telephone number. However, this does not happen very often because obtaining an unpublished number from directory assistance also requires that the number not also be designated as unlisted by the respondent or household.

The unpublished number disposition usually is considered a final disposition in telephone surveys of specifically named people (i.e., telephone surveys whose sample is a list of specific individuals), because it is highly unlikely that a researcher would be able to discover a telephone number for a sampled respondent or household if directory assistance and reverse directory techniques did not provide one. However, in a mixed-mode survey it might be possible to locate an address and to try to gather data by sending a mail questionnaire or an in-person interviewer to the address. As a result, researchers may consider the unpublished number disposition to be a temporary disposition in mixed-mode surveys.

Matthew Courser

See also Final Dispositions; Ineligible; Out of Order; Temporary Dispositions; Unlisted Household

Further Readings

American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (4th ed.). Lenexa, KS: Author.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Usability Testing

Although usability testing can apply to all types of products, for survey research, it can best be described as a method for measuring how well interviewers and respondents can use a computer-assisted interview such as a CAPI, CATI, CASI, or Web-based survey, for its intended purpose. It is important to separate usability testing from testing functionality, which focuses only on the proper operation of a computerized instrument (software and hardware), not the individual using the system. The purpose of usability testing is to determine whether or not the form being used to collect data helps or hinders a user's ability to deploy it.

In developing and designing survey instruments, researchers have always strived to ensure that data collection instruments are the best they can be through a variety of testing and evaluation methods put into place prior to data collection. Traditionally, cognitive interviewing and other cognitive methods have provided important tools for examining the thought processes that affect the quality of answers provided by survey respondents to survey questions. In addition, question appraisal systems are used to provide a structured, standardized instrument review methodology that assists a survey design expert in evaluating questions relative to the tasks they require of respondents, specifically with regard to how respondents understand and respond to survey questions. Focus groups can be used to obtain qualitative data that provide insight into the attitudes, perceptions, and opinions on a given topic or instrument. Although all of these efforts have long been important to understanding the way questions and the wording on a survey are perceived by respondents, the increased use of computer-assisted data collection has called for yet another form of testing instruments.

The general thought regarding computerized instruments is that they are easier on respondents and interviewers when compared with paper questionnaires. Pre-programmed skip patterns and automated progress through an instrument removes the time it takes to manually follow routing instructions, turn pages, and edit or calculate responses. But in practice, computer instruments can be more difficult to figure out than their paper counterparts because of complicated instructions, self-editing, navigational problems, and general layout. Usability testing can measure the time it takes to complete certain tasks in an instrument and whether or not these factors are contributing to increased respondent burden. Following the thought that burden is tied to stress or respondent fatigue, which could contribute to respondent attrition, identifying sources of burden and reducing them can contribute to improved response rates. In addition, usability testing can result in increased reliability and validity of survey instruments by examining features—such as error messages and other feedback, instructions, and placement of navigational features ("next buttons," etc.)—and assessing whether or not they help, confuse, encourage, or discourage respondents. The same examinations can also assist interviewers. Usability testing also can reveal how a computerized instrument affects the burden, emotions, and motivation of interviewers, which in turn, can have a positive impact on the quality of the data that they collect.

It is generally agreed that to properly conduct a high-quality usability test, a closed laboratory setting should be used. Many researchers use cognitive laboratories with common features such as one-way mirrors for observation to conduct usability testing. In addition, testing can be enhanced through the use of multiple cameras and recording devices. By using multiple cameras, researchers can capture users' hands on a computer keyboard as well as users' facial expressions. This practice is especially useful in allowing researchers to examine nonverbal cues that users may give, such as facial expressions or body language, that speak to burden or difficulties with a given task. By using microphones, researchers can record and analyze any comments that are made by users during testing. Devices such as scan converters or computers equipped with software allowing them to record images are useful for capturing images from a user's computer screen during testing. Video processors and editing equipment can also be used to capture images from all recording sources, synchronize them, and combine them so that the three images can either be viewed in real time or videotaped for later viewing, coding, and analysis.

Usability test subjects should reflect the population of interest. That said, it is important that usability tests focusing on interviewer-administered instruments are conducted with interviewers who are familiar with computer-assisted interviewing, while instruments to be used by the general public are tested among members of the general public. It is often recommended that tests are carried out with groups of 10 to 12 test subjects. Researchers will often prepare pre-scripted or mock interviews for test subjects to follow. By requiring test subjects to enter mock interview data, researchers can ensure that items of interest will be seen by test subjects. In addition, probes may be prepared in an effort to gain more detailed information about specific aspects of the instrument or task at hand that may not be mentioned spontaneously.

David James Roe

See also Cognitive Interviewing; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Contingency Question; Focus Group; Paper-and-Pencil Interviewing (PAPI); Questionnaire Design; Respondent Burden; Respondent Fatigue

Further Readings

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing* (Rev. ed.). Portland, OR: Intellect.

Hanson, S., & Couper, M. (2004). Usability testing to evaluate computer assisted instruments. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 337–360). New York: Wiley.

U.S. Bureau of the Census

The U.S. Bureau of the Census is a federal agency within the U.S. Department of Commerce. Most commonly known for conducting the enumeration of the U.S. population every 10 years, the bureau conducts activities that extend well beyond the decennial census. The largest federal statistical agency, the Census Bureau serves as the premier source of federal data on the U.S. population and economy. The Census Bureau is responsible for numerous surveys, including full censuses and sample surveys. The resulting data are used regularly not only by U.S. government officials but also by local governments, businesses, nonprofit organizations, and researchers in numerous disciplines.

Survey researchers in the United States are dependent on the accuracy of the decennial census and other Census Bureau surveys because they typically provide the population parameters (i.e., *universe estimates* [*UEs*]) to which other government, commercial, and academic survey data sets of the population often are weighted. Furthermore, because these other surveys are weighted to census data, the survey questionnaires used by these other organizations must use the same questionnaire wording used in census questionnaires so as to have equivalent variables for weighting purposes.

Census Bureau Surveys

The Census of Population and Housing (decennial census) is intended to count and collect information from all U.S. residents. First conducted in 1790, the decennial census is required by the U.S. Constitution for apportioning congressional representation. Conducting the census is a massive and multi-year effort. Major activities include finalizing content development, maintaining the Master Address File, printing and mailing paper questionnaires, conducting nonresponse follow-up, processing data, and disseminating results.

Modern population censuses are conducted largely through mailed paper questionnaires. In-person interviews occur when households do not have a usable mailing address or when residents do not return the mailed questionnaire (i.e., an in-person nonresponse follow-up). Recent decennial censuses have included short- and long-form questionnaires. With this design, long-form questionnaires included short-form questions as well as more detailed social, economic, and housing questions. Using statistical techniques, long forms were sent to a sample of U.S. residences (one in six for the year 2000), and results were projected to the total population or universe. Starting with the 2010 Census, only short-form content will be included on the decennial questionnaire. Long-form data will be collected through the American Community Survey, an ongoing survey with a sample of 3 million households per year.

The Census Bureau also conducts censuses of U.S. businesses and government entities. Like the population enumeration, these censuses are intended to fully cover all entities that meet the criteria for inclusion. The Economic Census and the Census of Governments are conducted every 5 years (years ending in the numerals 2 and 7). Like the decennial population census, federal code requires these censuses and mandates participation. The purpose of the Economic Census is to compile data on the U.S. economy through surveys of business establishments. The Census of Governments is intended to supply information on government organization, employment, and finance.

The Census Bureau has extensive expertise in survey research and often conducts surveys for other federal government agencies. Designed to fulfill many different federal data needs, these sample-based surveys vary in terms of methodology, frequency, and collection mode(s).

The Current Population Survey (CPS) is conducted by the Census Bureau for the Bureau of Labor Statistics. Conducted since the 1940s, the CPS is the main source for national labor statistics, including unemployment rates and earnings measures. The CPS is a large (50,000-household) national survey that is conducted over the telephone and in person. The CPS program includes an Annual Social and Economic Supplement, which includes more extensive demographic variables than the monthly surveys. Researchers often rely on the supplement for annually updated information during the years between decennial censuses. Because of its very high quality and sophisticated design, the CPS is often studied by survey researchers, statisticians, and other social scientists.

Other examples of Census Bureau surveys are the American Housing Survey and the Consumer Expenditure Survey. Conducted for the Department of Housing and Urban Development, the American Housing Survey collects data on housing quality, availability, and characteristics for the United States and certain metropolitan areas. The Consumer Expenditure Survey is a periodic consumer survey designed to collect information on purchasing behaviors. The Bureau of Labor Statistics uses data from the Consumer Expenditure Survey when calculating the Consumer Price Index, a measure of inflation that tracks changes in the prices of products and services.

Applications for Survey Researchers

Survey researchers and social scientists rely on federal census results and sample-based estimates as valuable and comprehensive sources of information about the U.S. population and economy. Census data are subject to not only sampling error but also other survey errors, such as incomplete coverage and nonresponse. After the decennial census, the bureau publishes estimates of the net undercount or overcount as well as various nonresponse rates. Compared to other surveys, the error rates for the census are often very low. The sample-based surveys are subject to both sampling and nonsampling error. But the very large sample sizes and benefits of conducting ongoing surveys (such as professionally trained interviewers) often yield high-quality results.

Federal census results and sample-based estimates often serve as population or household controls for other sample-based surveys. The data can be used in various sampling phases such as design, stratification, and selection. Census data are commonly used as UEs in sample weighting adjustments, which reduce bias by accounting for differences in characteristics between a sample and the universe it is designed to represent.

Through extensive research on survey methods and statistical techniques, the Census Bureau strives to improve its own data collection and contribute to a wider knowledge base. Census Bureau reports cover topics such as survey design, questionnaire effects, and interviewing procedures. The resulting body of research provides very useful information to survey research professionals and students. Census Bureau technical reports and working papers are often accessible through federal Web sites or by request.

Christine Pierce

See also American Community Survey (ACS); Bureau of Labor Statistics (BLS); Census; Current Population Survey (CPS); Sample; Survey

Further Readings

Anderson, M. J. (1988). *The American census: A social history.* New Haven, CT: Yale University Press.

Anderson, M. J., & Fienberg, S. E. (1999). *Who counts? The politics of census-taking in contemporary America.* New York: Russell Sage Foundation.

Bureau of Labor Statistics. (2002, March). *Design and methodology* (Current Population Survey Tech. Paper No. 63RV). Retrieved January 10, 2007, from http://www.census.gov/prod/2002pubs/tp63rv.pdf

Executive Office of the President, Office of Management and Budget. (2006, September). *Statistical programs of the United States government: Fiscal year 2007.* Washington, DC: Author. Retrieved January 15, 2007, from http://www.whitehouse.gov/omb/inforeg/07statprog.pdf

Prewitt, K. (2003). *Politics and science in census taking.* New York: Russell Sage Foundation.

U.S. Census Bureau. (2008). *Statistical Abstract of the United States.* Washington, DC: U.S. Department of Commerce, Bureau of the Census, Government Printing Office. Retrieved April 24, 2008, from http://www.census.gov/compendia/statab/

U.S. Department of Commerce, Economics and Statistics Administration, & U.S. Census Bureau. (2007, June). *U.S. Census Bureau Strategic Plan FY 2007–2012.* Retrieved April 24, 2008, from http://www.census.gov/main/www/strategicplan/strategicplan.pdf

V

VALIDATION

Within survey research, *validation* has more than one meaning, but this entry focuses on the concept of validating a completed interview, in particular one completed via paper and pencil, as soon as possible after it has been completed—that is, making certain that (a) all questions that should have been asked and answered were done so, and (b) answers that were written in for open-ended questions are legible, understandable, and relevant to the question being asked. The advantages of doing this type of validation as soon as possible after interviews are completed are three-fold. First, it allows for mistakes to be corrected soon after they have been made at a time when an interviewer's memory of the interview is still relatively fresh. Second, especially early in the field period of a survey, it helps to identify the need for retraining all or some interviewers if consistent problems are detected by the person doing the validation. Third, early in the survey period, it allows the researchers to modify their questionnaire if that is necessary—for example, fix a skip pattern that is not working as intended—before the problem negatively affects too many of the completed cases that otherwise may need to be reinterviewed.

With the advent of computer-assisted interviewing, whether the questionnaire is administered by an interviewer or is self-administered by the respondent, the need for manual validation of completed interviews is greatly reduced but not necessarily eliminated. It is reduced because the logic that is built into the computer interviewing software will not allow many of the problems to occur that are often present in paper-and-pencil interviewing (PAPI). These include erroneously skipped questions that should have been asked of a particular respondent and other questions that were asked when they should have been skipped. Problems like these in PAPI interviewing can be caught through manual validation and quickly corrected—whether they indicate the need to change the skipping instructions for certain contingency questions or the need to retrain particular interviewers.

Furthermore, the need for manual validation is not eliminated entirely by computer-assisted interviewing because the answers to open-ended questions cannot be checked for quality via computer software. Imagine a researcher who did not have open-ended answers validated until after all data were gathered, only to find significant problems in what was recorded by interviewers or respondents. Such problems could result from many reasons, including an open-ended question that was not well understood by most respondents. If this validation did not happen until all the data were gathered, the researcher may need to scrap this variable entirely. Instead, by having open-ended responses manually validated in the early stages of a survey, the researchers can catch such problems before they affect too many respondents. This will help ensure data quality and ultimately avoid disasters where those data for an entire variable are worthless.

Paul J. Lavrakas

See also Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Contingency Question; Open-Ended Question; Quality Control; Verification

Further Readings

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

VALIDITY

The word *validity* is primarily a measurement term, having to do with the relevance of a measuring instrument for a particular purpose, but it has been broadened to apply to an entire study. A research investigation is said to have *internal validity* if there are valid causal implications and is said to have *external validity* if the results are generalizable.

As far as measurement is concerned, the most important property of a measuring instrument is the extent to which it has been validated with respect to some gold standard whose validity has been assumed to be taken for granted. For example, if scores on a test of mathematical aptitude (the instrument to be validated) correlate highly with scores on a subsequent test of mathematical achievement (the gold standard), all is well, and the aptitude test would be regarded as valid.

In the early 20th century, the methodological literature referred to three kinds of validity: (1) *content validity* (expert judgment, i.e., postulation of the gold standard itself), (2) *criterion-related validity* (the type of validity mentioned in the previous paragraph), and (3) *construct validity* (the extent to which the scores obtained using a particular measuring instrument agreed with theoretical expectations). There were also subtypes of criterion-related validity (*concurrent* and *predictive*) and construct validity (*convergent* and *discriminant*), but more recently the general label "construct validity" not only has become more popular but also has been alleged to include content validity and criterion-related validity as well.

Connection Between Reliability and Validity

Reliability is concerned with the consistency of results whether or not those results are valid. It is easy to imagine a situation in which ratings given to various contestants (in a figure skating event, for example) are consistent (reliable) from one judge to another, but that all the ratings are wrong due to the personal biases of the judges.

There is a way to investigate the validity of an instrument with respect to its reliability. Suppose that a survey researcher were interested in pursuing further the relationship between performance on an aptitude test (the predictor) and performance on an achievement test (the criterion with respect to which the predictor instrument is to be validated) in which a correlation between the two of .54 has been obtained. The researcher would like to estimate what the correlation would be if it were based upon true scores rather than observed scores. There is a formula called *the correction for attenuation* that can be used for that purpose. The obtained correlation is divided by the product of the square roots of the estimates of the respective reliability coefficients. If the reliability coefficient for the aptitude test were estimated to be .64 and the reliability coefficient for the achievement test were estimated to be .81, application of that formula would yield a value of .75, which is considerably higher than the .54. That makes sense, because the true correlation has been attenuated (i.e., reduced) by the less-than-perfect reliabilities of the two tests.

Validity of Instrument Versus Validity of Scores Obtained With Instrument

Just as for reliability, it is somewhat controversial whether researchers should refer to the validity of a measuring instrument or to the validity of the scores obtained with the instrument. Some authors even go so far as to insist that any investigation of the validity of a measuring instrument should address the consequences of any actions taken on the basis of such scores.

Thomas R. Knapp

See also Construct Validity; External Validity; Internal Validity; Reliability; True Value

Further Readings

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, UK: Cambridge University Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-017). Beverly Hills, CA: Sage.

Litwin, M. S. (1995). *Survey kit: Vol. 7. How to measure survey reliability and validity*. Thousand Oaks, CA: Sage.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

VARIABLE

A variable is something that varies in value, as opposed to a constant (such as the number 2), which always has the same value. These are observable features of something that can take on several different values or can be put into several discrete categories. For example, respondents' scores on an index are variables because they have different values, and religion can be considered a variable because there are multiple categories. Scientists are sometimes interested in determining the values of constants, such as π, the ratio of the area of a circle to its squared radius. However, survey research involves the study of variables rather than constants.

A quantity X is a *random variable* if, for every number a, there is a probability p that X is less than or equal to a. A *discrete random variable* is one that attains only certain values, such as the number of children one has. By contrast, a *continuous random variable* is one that can take on any value within a range, such as a person's income (measured in the smallest possible fractions of a dollar).

Data analysis often involves hypotheses regarding the relationships between variables, such as "If X increases in value, then Y tends to increase (or decrease) in value." Such hypotheses involve relationships between *latent variables*, which are abstract concepts. These concepts have to be operationalized into *manifest variables* that can be measured in actual research. In surveys, this operationalization involves either using one question to tap the concept or combining several questions into an index that measures the concept.

A basic distinction in statistical analysis is between the dependent variable that the researcher is trying to explain and the independent variable that serves as a predictor of the dependent variable. In regression analysis, for example, the dependent variable is the Y variable on the left-hand side of the regression equation $Y = a + bX$, whereas X is an independent variable on the right-hand side of the equation.

The starting point in survey analysis is often looking at the distribution of the variables of interest, one at a time, including calculating appropriate univariate statistics such as percentage distributions. The changes in that variable over time can then be examined in a time-series analysis. Univariate analysis is usually just the jumping-off point for bivariate or multivariate analysis. For example, in survey experiments, the researcher examines the extent to which experimental manipulations in the survey (such as alternative wordings of a question) affect the variance in the variable of interest.

Herbert F. Weisberg

See also Dependent Variable; Experimental Design; Independent Variable; Null Hypothesis; Regression Analysis

Further Readings

Lewis-Beck, M. S. (1995). *Data analysis: An introduction.* Thousand Oaks, CA: Sage.

VARIANCE

Variance, or dispersion, roughly refers to the degree of scatter or variability among a collection of observations. For example, in a survey regarding the effectiveness of a political leader, ratings from individuals will differ. In a survey dealing with reading ability among children, the expectation is that children will differ. Even in the physical sciences, measurements might differ from one occasion to the next because of the imprecision of the instruments used. In a very real sense, it is this variance that motivates interest in statistical techniques.

A basic issue that researchers face is deciding how variation should be measured when trying to characterize a population of individuals or things. That is, if all individuals of interest could be measured, how should the variation among these individuals be characterized? Such measures are population measures of variation. A related issue is deciding how to estimate a population measure of variation based on a sample of individuals.

Choosing a measure of dispersion is a complex issue that has seen many advances during the past 30 years, and more than 150 measures of dispersion have been proposed. The choice depends in part on the goal of the investigator, with the optimal choice often changing drastically depending on what an investigator wants to know or do. Although most of these measures seem to have little practical value, at least five or six play an important and useful role.

Certainly the best-known measure of dispersion is the *population variance*, which is typically written as σ^2. It is the average (or expected) value of the squared difference between an observation and the population mean. That is, if all individuals in a population could be measured (as in a complete census), the average of their responses is called the population mean, μ, and if for every observation the squared difference between it and μ were computed, the average of these squared values is σ^2. In more formal terms, $\sigma^2 = E(X - \mu)^2$, where X is any observation the investigator might make and E stands for expected value. The (positive) square root of the variance, σ, is called the (population) standard deviation.

Based on a simple random sample of n individuals, if the investigator observes the values $X_1, \ldots, X_n$, the usual estimate of σ^2 is the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$ is the sample mean.

For some purposes, the use of the standard deviation stems from the fundamental result that the probability of an observation being within some specified distance from the mean, as measured by σ, is completely determined under normality. For example, the probability that an observation is within one standard deviation of the mean is .68, and the probability of being within two standard deviations is .954. These properties have led to a commonly used measure of effect size (a measure intended to characterize the extent to which two groups differ) as well as a frequently employed rule for detecting outliers (unusually large or small values). Shortly after a seminal paper by J. W. Tukey in 1960, it was realized that even very small departures from normality can alter these properties substantially, resulting in practical problems that commonly occur.

Consider, for example, the two distributions shown in Figure 1. One is a standard normal, meaning it has a mean of zero (0) and a variance of one (1). The other is not a normal distribution, but rather something called a mixed or contaminated normal. That is, two normal distributions, both having the same mean but different variances, are mixed together. The important point is that despite the obvious similarity between the two distributions, their variances differ substantially: The normal has variance 1, but the mixed normal has variance 10.9. This illustrates the general principle that small changes in the tails of a distribution can substantially affect σ.

Now consider the property that under normality, the probability that an observation is within one standard deviation of the means is .68. Will this property hold under nonnormality? In some cases it is approximately true, but in other situations this is no longer the case, even under small departures from normality (as measured by any of several commonly used metrics by statisticians). For the mixed normal distribution in Figure 1, this probability exceeds .999.

The *sample variance* reflects a similar property: It is very sensitive to outliers. In some situations this can be beneficial, but for a general class of problems it is detrimental. For example, a commonly used rule is to declare the value X an outlier if it is more than

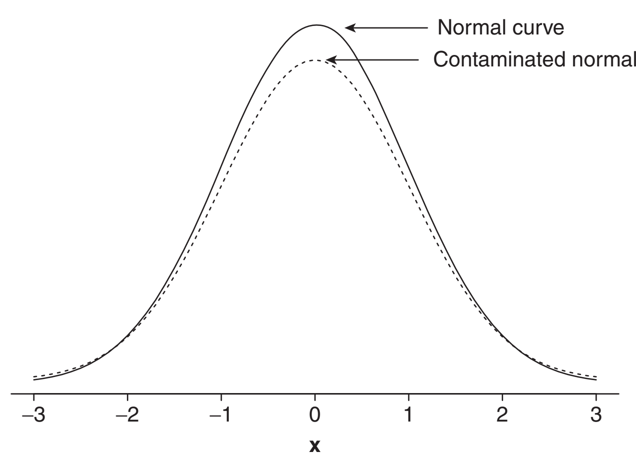

Figure 1 Variances of a normal distribution and a mixed normal distribution

two (sample) standard deviations from the mean, the idea being that under normality such a value is quite unusual from a probabilistic point of view. That is, declare X an outlier if

$$\frac{X - \bar{X}}{s} > 2.$$

This rule suffers from *masking*, however, meaning that even obvious outliers are not detected due to the sensitivity of s to outliers. Consider, for example, the values 2, 2, 3, 3, 3, 4, 4, 4, 100,000, 100,000. Then $\bar{X} = 20002.5$, $s = 42162.38$, and $(100{,}000 - 20002.5)/s = 1.9$. So the value 100,000 would not be declared an outlier based on the rule just given, yet it is clearly unusual relative to the other values.

For detecting outliers, two alternative measures of dispersion are typically used. The first is called the *median absolute deviation (MAD)* statistic. It is the median of the n values $|X_1 - M|, \ldots, |X_n - M|$, where M is the usual median of $X_1, \ldots, X_n$. Now X is declared an outlier if

$$\frac{|X - M|}{MAD/.6745} > 2.24.$$

The other is the *interquartile range*, which is just the difference between the upper and lower quartiles. (This latter measure of dispersion is used by *boxplot rules* for detecting outliers.) An important feature of both of these measures of dispersion is that they are insensitive to extreme values; they reflect the variation of the centrally located observations, which is a desirable property when detecting outliers with the goal of avoiding masking.

Another approach when selecting a measure of dispersion is to search for an estimator that has a relatively low standard error over a reasonably broad range of situations. In particular, it should compete well with s under normality, but it is desirable to maintain a low standard error under nonnormality as well. So, for example, if it is desired to compare two groups in terms of dispersion, the goal is to maintain high power regardless of whether sampling is from a normal distribution or from the mixed normal distribution in Figure 1.

Two measures of dispersion that satisfy this goal are called the *percentage bend midvariance* and the *biweight midvariance.* The tedious computational details are not given here, but they can be found in the further readings, along with easy-to-use software. In terms of achieving a relatively low standard error, the sample standard deviation, s, competes well with these two alternative measures under normality, but for nonnormal distributions, s can perform rather poorly.

Effect Size

The variance has played a role in a variety of other situations, stemming in part from properties enjoyed by the normal distribution. One of these is a commonly used measure of *effect size* for characterizing how two groups differ:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma_p},$$

where μ_1 and μ_2 are the population means, and where, by assumption, the two groups have equal standard deviations, σ_p. The idea is that if $\Delta = 1$, for example, the difference between the means is one standard deviation, which provides perspective on how groups differ under normality. But concerns have been raised about this particular approach because under nonnormality it can mask a relatively large difference.

Standard Errors

The notion of variation extends to sampling distributions in the following manner. Imagine a study based on n observations resulting in a simple random sample mean, say $\bar{X}$. Now imagine the study is repeated many times (in theory, infinitely many times) yielding the sample means $\bar{X}_1, \bar{X}_2, \ldots$, with each sample mean again based on n observations. The variance of these sample means is called the *squared standard error of the sample mean*, which is known to be

$$E(\bar{X} - \mu)^2 = \sigma^2/n$$

under random sampling. That is, it is the variance of these sample means. Certainly the most useful and important role played by the variance is making inferences about the population mean based on the sample mean. The reason is that the standard error of the sample mean, $\sigma/\sqrt{n}$, suggests how inferences about μ should be made assuming normality, but the details go beyond the scope of this entry.

Winsorized Variance

Another measure of dispersion that has taken on an increasing importance in recent years is the *Winsorized variance*; it plays a central role when comparing groups based on *trimmed means*. It is computed as follows. Consider any γ such that $0 \le \gamma < 1$. Let $g = \gamma n$, rounded down to the nearest integer. So if $\gamma n = 9.8$, say, $g = 9$. Computing a γ-trimmed mean refers to removing the g smallest and g largest observations and averaging those that remain. Winsorizing n observations refers to setting the g smallest values to the smallest value not trimmed and simultaneously setting the g largest values equal to the largest value not trimmed. The sample variance, based on the Winsorized values, is called the *Winsorized variance*. Both theory and simulation studies indicate that trimming can reduce problems associated with means due to nonnormality, particularly when sample sizes are small. Although not intuitive, theory indicates that the squared standard error of a trimmed mean is related to the Winsorized variance, and so the Winsorized variance has played a role when making inferences about a population trimmed mean. Also, it has been found to be useful when searching for robust analogs of Pearson's correlation.

Rand R. Wilcox

See also Mean; Outliers; Standard Deviation; Standard Error; Variance Estimation

Further Readings

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

VARIANCE ESTIMATION

Variance refers to the degree of variability (dispersion) among a collection of observations. Although estimation of the size of the variance in a distribution of numbers often is a complex process, it is an extremely important endeavor for survey researchers, as it helps make valid inferences of population parameters.

Standard variance estimation formulas for simple random sampling, stratified sampling, and cluster sampling can be found in essentially any survey sampling textbook, such as those by William G. Cochran or Leslie Kish. However, most large survey samples use combinations of unequal probabilities, stratification, and clustering. Often, survey weights are used in estimation that account for unequal probabilities, nonresponse, and post-stratification. These are usually ratio estimates with the numerator being the weighted average and the denominator being the sum of the weights. Both the numerator and denominator are random variables. However, textbook variance formulas are not sufficient for these survey samples and estimation problems. Specialized variance estimation software packages have been required until only recently, but now general-purpose statistical analysis programs have started to include the special variance estimation techniques needed to correctly calculate variances for complex sample surveys.

Without correct variance estimates, users are unable to make valid inferences concerning population parameters. Most complex sample surveys have larger standard errors than do simple random samples. If inference is done using standard errors from simple random samples, the standard errors would be too small and any statistical procedures would be too liberal (e.g., p-values or confidence intervals would be too small, and test statistics would be too large).

Calculation Methods

The two most popular classes in the calculation of correct variance estimates for complex sample surveys are *replicate methods* for variance estimation and *Taylor series linearization methods*. A third class of techniques that are sometimes used are *generalized variance functions*.

Replicate methods for variance compute multiple estimates in a systematic way and use the variability in these estimates to estimate the variance of the full-sample estimator. The simplest replicate method for variance is the *method of random groups*. The method of random groups was originally designed for interpenetrating samples or samples that are multiple repetitions (e.g., 10) of the same sampling strategy. Each of these repetitions is a random group, from which an estimate is derived. The overall estimator is the average of these estimates, and the estimated variance of the estimator is the sampling variance of the estimators. Of course, this technique can be used for any complex sample survey by separating the sample into subsamples that are as equivalent as possible (e.g., by sorting by design variables and using systematic sampling to divide into random groups). This simplest of replicate methods for variance is simple enough to do by hand, but is not as robust as more modern replication-based methods that can now be easily calculated by computers.

Balanced repeated replication (BRR), also known as *balanced half-samples*, was originally conceived for use when two primary sampling units (PSUs) are selected from each stratum. A half-sample then consists of all cases from exactly one primary sampling unit from each stratum (with each weight doubled). Balanced half-sampling uses an orthogonal set of half-samples as specified by a Hadamard matrix. The variability of the half-sample estimates is taken as an estimate of the variance of the full-sample estimator.

Jackknife variance calculation, also called *Jackknife repeated replication (JRR)*, works in a similar way to BRR. The JRR method creates a series of replicate estimates by removing only one primary sampling unit from only one stratum at a time (doubling the weight for the stratum's other primary sampling unit).

For a simple example, let us use two strata (1 and 2) with two PSUs in each (1a, 1b, 2a, and 2b). BRR could form estimates by using only 1a and 2a, only 1b and 2b, only 1a and 2b, and only 1b and 2a (each orthogonal combination of PSUs). JRR drops only one PSU at a time and forms estimates by using 1b (doubly weighted), 2a, and 2b; or 1a (doubly weighted), 2a, and 2b; or 1a, 1b, and 2a (doubly weighted); or 1a, 1b, and 2b (doubly weighted).

Even if a complex sample survey does not have exactly two PSUs in each stratum, the BRR and JRR techniques can be used by defining pseudo-strata and pseudo-PSUs so that there are two pseudo-PSUs in each pseudo-stratum. Of course, these methods work only as well as these definitions work.

A more general replication method for variance is *bootstrapping*. The idea of bootstrapping is that, from the full original sample drawn, the researcher repeatedly draws simple random samples of the same sample size with replacement. With each draw, different observations will be duplicated, and these differences will result in different estimates. The variance of these estimates is the variance estimate.

Taylor series linearization methods work by approximating the estimator of the population parameter of interest by a linear function of the observations. These approximations rely on the validity of Taylor series or binomial series expansions. An estimator of the variance of the approximation is then used as an estimator of the variance of the estimator itself. Generally, first-order approximations are used (and are adequate), but second- and higher-order approximations are possible.

A third, but less commonly used, class of variance estimation methods is a *generalized variance function (GVF).* Generalized variance functions are tables or formulas prepared by statistical experts as guidance for analysts who cannot directly calculate correct standard errors. The experts calculate variance estimates for as many variables as is practical and use the direct estimates as input into a regression model. The most common model states that the squared coefficient of variance is a linear function of the inverse of its expectation:

$$\frac{Var(\hat{X})}{X^2} = a + \frac{b}{X}.$$

Regression estimates for a and b can be done separately for multiple subgroups, or domains. Analysts can then input their estimate (X) and the two parameters a and b to calculate an estimated variance. GVFs are still being used, but as variance estimation tools become easier to use, the use of GVFs will continue to decline.

Choosing a method from among those described in this entry is an important and difficult decision. Software has improved so that those who can use standard statistical software now have the ability to calculate correct variance estimates. Empirical results from comparisons between the replicate methods for variance estimation and Taylor series linearization methods show only subtle differences (there is no difference asymptotically). Therefore, often the best solution is for researchers to use software with which they are already familiar. There seems a slight preference in the literature for the balanced repeated replication method on accuracy. However, Taylor series linearization methods are easier to set up and use. Jackknife variance calculations are less available through software, and bootstrapping software is very limited. The random groups method has been superseded by the other three replicate methods for variance estimation; its main advantage is its simplicity. Generalized variance functions are useful mostly for publication so that analysts can calculate reasonable (and stable) variance estimates without using the other statistical techniques; individual estimates of variance using the functions will be less accurate than other statistical methods. Finally, it is generally not acceptable to use simple random sample formulas to calculate variance estimates for complex sample surveys.

Software for Variance Estimation

Standard statistical software packages such as SAS, STATA, and Statistical Package for the Social Sciences (SPSS) used to only be able to calculate standard errors assuming a simple random sample. Standard routines in these packages still make this assumption. Such standard errors tend to be too small for a cluster sample design. However, these packages have recently added capability through new procedures or modules for calculating design-corrected standard errors for complex sample surveys.

Nevertheless, specialized variance estimation programs such as SUDAAN and WesVar are superior for variance estimation because they have wider sets of analyses for which correct variance estimation can be done, and the implementation is more robust after having such a head start on the general-use statistical software packages.

Steven Pedlow

See also Balanced Repeated Replication (BRR); Bootstrapping; Clustering; Cluster Sample; Cochran, W. G.; Complex Sample Surveys; Jackknife Variance Estimation; Kish, Leslie; Primary Sampling Unit; Replicate Methods for Variance Estimation; Sampling Variance; SAS; Simple Random Sample; Standard Error; Stata; Statistical Package for the Social Sciences (SPSS); Stratified Sampling; SUDAAN; Taylor Series Linearization; WesVar

Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Kish, L. (1995). *Survey sampling.* New York: Wiley.
Wolter, K. (1985). *Introduction to variance estimation.* New York: Springer.

VERBATIM RESPONSES

A verbatim response refers to what an interviewer records as an answer to an open-ended question when writing down the *exact words* spoken by the respondent. Open-ended questions are those that do not provide a respondent with predetermined response choices and instead allow, expect, and encourage a respondent to answer in his or her own words.

Sometimes open-ended questions have "precoded" response categories for interviewers to use, but these are not read to a respondent. In these instances interviewers are trained to listen to the answer a respondent gives and then "fit" it into one of the precoded choices. However, this is not always possible to do for all given responses, so even with precoded open-ended questions, some verbatim responses that do not fit into any one of the precoded categories must be written down by the interviewer. Most open-ended questions do not have precoded response choices for interviewers to code; thus in most cases with an open-ended question, the interviewer must write down what the respondent says.

In some cases a researcher may allow interviewers to summarize the gist of what the respondent says in response to an open-ended question. However, in other cases it is important that the exact words spoken by the respondent are recorded by the interviewer. It is in these cases that interviewers are trained and expected to write down the complete verbatim response. Interviewers are trained to "slow down" respondents who are speaking too fast in answering an open-ended question for which a verbatim response must be recorded by the interviewer. Interviewers also are trained to use probing techniques to get more detailed replies to open-end questions.

After data collection has ended, the verbatim responses that interviewers have recorded often are coded (similar to content analysis) so that they then can be analyzed via quantitative statistical techniques. This is a labor-intensive and expensive process, if done reliably. Verbatim responses also are used in survey reports to help illustrate the statistical findings gleaned from coded open-ended variables by putting a "human face" to them. This is done without identifying the respondents whose verbatim responses are used, so as not to violate the confidentiality pledge given to the respondents.

Paul J. Lavrakas

See also Content Analysis; Open-Ended Question; Precoded Question; Probing; Standardized Survey Interviewing

Further Readings

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing.* Newbury Park, CA: Sage.

VERIFICATION

Within survey research, one of the meanings of *verification* refers to efforts that are made to confirm that a telephone or in-person interview was actually completed with a particular respondent. Within this context, verification is one of the techniques used to guard against interviewer falsification. Such verification often is done via telephone even if the interview was done in person. In most cases a small number (e.g., less than five) of randomly selected completed interviews of all interviewers are verified by having a person in a supervisory position contact the respondents and verify that they were, in fact, interviewed by the interviewer. In other instances, interviewers who may be under observation for previous interviewing infractions may have a greater number of their interviews verified. Oftentimes, interviewers are instructed to tell respondents at the end of an interview that someone may contact them to verify that the interview was completed.

Other meanings of verification, within the context of survey research, concern efforts that researchers sometimes make to verify whether a respondent's answer to a key variable is accurate. These procedures include so-called *record checks* in which the researcher gains access to an external database (such as those assembled by various government agencies) that can be used to officially verify whether the answer given by a respondent is correct. For example, in a survey conducted of voting behavior, if respondents are asked in a survey whether they voted in the past election, then researchers could go to public election records to learn if a given respondent did or did not vote in the last election, thereby verifying the answer given in the survey. However, in this entry, it is the former meaning of the term *verification* that is addressed.

In theory, informing interviewers that some of their completed interviews will be verified is assumed to motivate them against any falsification. However, little empirical work has been done to test this assumption. Nevertheless, many clients expect this will be done, and as such, verification is a process that many survey organizations build into their contracts. Of note, verification should not be confused with *interviewer monitoring*, which is a process used in real time to observe the behavior of interviewers as they interact with respondents in (a) gaining cooperation from the respondents and (b) gathering data from them.

Verification serves additional purposes beyond its primary purpose of confirming whether or not a given interview was conducted. In addition, a systematic and routine verification process provides information that a survey organization can use to evaluate the job performance of individual interviewers. For example, when respondents are contacted to verify a completed interview, they also can be asked a few additional questions, at very little additional cost to the survey organization, to help evaluate the quality of the interviewer's work. A system of routine verification can also serve as a check on the quality of the interviewer monitoring system a survey organization has instituted, as verification is a fail-safe approach to detecting falsification, which ideally should be detected via on-the-job monitoring.

Paul J. Lavrakas

See also Falsification; Interviewer Monitoring; Record Check; Reverse Record Check; Validation

Further Readings

Braun, J. R., & La Faro, D. (1968). Effectiveness of the California Study Methods Survey verification key in detecting faking. *Journal of Educational Measurement*, *5*(2), 167–168.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

VIDEO COMPUTER-ASSISTED SELF-INTERVIEWING (VCASI)

Video computer-assisted self-interviewing (VCASI), sometimes referred to as audio-visual CASI (AVCASI), involves administration of an electronic self-administered questionnaire, which includes prerecorded video clips of an interviewer asking survey questions. This mode of data collection is an extension of audio computer-assisted self-interviewing (ACASI). Responses are recorded either by keyboard, mouse, or touch screen data entry. This technique is typically used in two settings. First is in conjunction with a face-to-face computer-assisted interview, where part of the interview is conducted by the interviewer and part is left to the respondent using the computer in a self-administered mode. Second is in a completely self-administered setting where the CASI questionnaire runs on a stand-alone terminal or kiosk. In an interviewer-administered setting, the respondent uses headphones for complete privacy. In a private kiosk setting the questions might be delivered to the respondent using built-in computer speakers.

VCASI was developed to collect information on sensitive topics, such as illegal drug use, sexual preference, criminal activity, in as valid and reliable a manner as possible. When responding to sensitive questions in the presence of a live interviewer or with other persons present, respondents are sometimes reluctant to answer certain questions completely and honestly. Instead, they tend to provide partial responses, socially desirable responses, or refuse to answer altogether. To overcome such resistance, interviewers equip respondents with headphones, which prevent others from overhearing the questions. The respondent then enters his or her response directly into the computer, thereby ensuring greater confidentiality. This approach is assumed to be less prone to social desirability bias and other response errors related to the presence of an interviewer or third parties. Unlike ACASI, in which the respondent typically hears only the questions being asked and sees text or a visual representation of the question on the computer screen, with VCASI the video of the interviewer is thought to mimic some of the benefits associated with face-to-face interviewing, such as personal interaction.

In addition, VCASI helps address the issue of illiteracy among some respondents, which can be a problem with some self-administered survey modes (such as mail surveys). Because the questions are asked in a video format, respondents do not need to be able to read the questions. They do, however, need to be able to enter the correct response category, which typically is an easier task requiring a lower level of reading competency.

VCASI also provides standardization of interviewer behaviors. Because the interviewer is prerecorded, there is no interviewer variation per se, such as interviewer mood or voice inflection, which might affect how a person responds. Conversely, interviewers cannot react to personal characteristics of the respondent, the interview setting, or to the responses provided and thereby bias the data. In addition, because skip instructions and branching, as well as the sequence of the questions, are predefined and programmed into the VCASI system, no accidental or intentional omission of questions can occur, and all questions are administered exactly as worded. VCASI is not suitable, however, for complicated concepts and questions that require probing or clarification. Also, lengthy questionnaires that may require motivational efforts from an interviewer to maintain the respondent's interest in the survey are not good candidates for this approach.

In a kiosk setting (such as in a museum or shopping mall) where third parties cannot be excluded from the interview setting, the increased privacy provided by VCASI is less of a benefit. In such settings, however, it is a cost-efficient method of data collection, because no interviewers are needed and data entry is conducted by respondents at little to no cost, other than the fixed costs associated with system development, programming, and maintenance.

Marek Fuchs

See also Audio Computer-Assisted Self-Interviewing (ACASI); Interviewer Variance; Sensitive Topics; Social Desirability; Underreporting

Further Readings

Katz, L. M., Cumming, P. D., & Wallace, E. L. (2007). Computer-based blood donor screening: A status report. *Transfusion Medicine Reviews*, *21*(1), 13–25.

Kissinger, P., Rice, J., Farley, T., Trim, S., Jewitt, K., Margavio, V., et al. (1999). Application of computer-assisted interviews to sexual behaviour research. *American Journal of Epidemiology*, *149*(10), 950–954.

O'Reilly, J. M., Hubbard, M. L., Lessler, J. T., Biemer, P. P., & Turner, C. F. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technologies for data collection. *Journal of Official Statistics*, *10*(2), 197–214.

VIDEOPHONE INTERVIEWING

Videophone interviewing enables researchers to closely reproduce face-to-face interviewing without requiring interviewers to physically enter respondent homes. A videophone enables real-time video and audio communication between two parties. There are two types of videophones: landline and mobile. Landline videophones are telephones with an LCD screen and a video camera. Mobile videophones are similar to landline phones with the exception of having a smaller screen. To transmit real-time video conversation, mobile phones should have 3G service—a wireless technology capable of transmitting a large amount of data.

Because videophone interviews closely mimic face-to-face interviews, the interaction between interviewer and respondent also shares the benefits of face-to-face interviewing. Specifically, the benefit of observing nonverbal communication, of allowing a slower pace than phone interviews, and of overcoming problems of respondent distraction are some examples. Because interviewers cannot observe the nonverbal communication of the respondent in a regular phone survey, it is more difficult to spot comprehension problems and to address issues of respondent motivation. Furthermore, interviewers cannot send nonverbal motivational cues or encouragement. Silences are also more difficult to handle on a regular telephone than face-to-face. In a videophone interview, the pace of the conversation can be reduced in comparison to a telephone interview. Breaking off would be more difficult than in a telephone survey. Lastly, multi-tasking, such as doing something else while being on the phone, can be discouraged in a videophone interview, as the norm of politeness is likely relevant. In other words, a videophone interview should elicit higher-quality data when compared to telephone interviews.

The videophone enables researchers to send text and multi-media material to the respondent. For example, show cards can be used when properly formatted for the small screen. In addition, still pictures, audio, and video can be sent to the respondent as is done nowadays with Web surveys.

Videophone interviewing is likely the most similar method to face-to-face interviewing, an interviewing technique that is still considered the most flexible and reliable, allowing for complex and long interviews, the reduction of missing data, and the increase response rates. Videophone interviewing, however, will potentially reintroduce some interviewer effects common in face-to-face interviews (e.g., social desirability bias).

Videophone interviews are still in the experimental phase. They are used in both the medical field and in psychological fields to communicate with and diagnose patients. Landline videophone penetration rates are extremely low or nonexistent in many countries, thus rendering landline videophone interviewing not feasible for surveys of the general population. On the other hand, the percentage of mobile phones equipped with videophone capabilities is growing in many industrialized countries, making videophone interviews a future possibility for survey researchers.

Mario Callegaro

See also Face-to-Face Interviewing; Show Card; Social Desirability; Telephone Surveys; Video Computer-Assisted Self-Interviewing (VCASI); Web Survey

Further Readings

Couper, M. P., Tourangeau, R., & Steiger, D. M. (2001, November). *Social presence in Web surveys.* Paper presented at the annual Federal Committee on Survey Methodology conference, Arlington, VA. Retrieved April 24, 2008, from http://www.fcsm.gov/01papers/Tourangeau.pdf

Demiris, G., Parker Oliver, D., & Courtney, K. (2006). A study of the suitability of videophone for psychometric assessment. *Behaviour & Information Technology*, *25*, 233–237.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*, 79–125.

VIGNETTE QUESTION

A vignette is a sort of "illustration" in words. In survey research, a vignette question describes an event, happening, circumstance, or other scenario, the wording of which often is experimentally controlled by the researcher and at least one of the different versions of the vignette is randomly assigned to different subsets of respondents.

For example, imagine a vignette that describes a hypothetical crime that was committed, and the respondents are asked closed-ended questions to rate how serious they consider the crime to be and what sentence a judge should assign to the perpetrator. The researcher could experimentally alter the wording of the vignette by varying six independent variables: the gender, race, and age of both the victim and perpetrator. If two ages (e.g., 16 and 53), three races (Asian, black, and white) and both genders (female and male) were varied for both the victim and perpetrator, the resulting design—a $2 \times 3 \times 2 \times 2$ experiment—would yield 24 different versions of the vignette. In the typical use of this approach, one of these 24 versions would be randomly assigned to each respondent. This would allow the researcher to test for three main effects (age, gender, race), three two-way interaction effects, and one three-way interaction effect.

The experimental use of vignettes has been enhanced greatly with the advent of computer-assisted survey data collection such as CAPI, CASI, and CATI. No longer do researchers need to print hard copies of multiple versions of a questionnaire to deploy randomly to respondents. Instead, the computer software randomly serves up the version or versions of the vignette to any one respondent, thus balancing the allocation of the different versions easily and correctly across the entire sample.

Another way vignettes are used in survey research is illustrated as follows. A vignette question could describe a protagonist (or group of protagonists) faced with a realistic situation pertaining to the construct under consideration. The respondent is asked to make a judgment about the protagonist, the situation, or the correct course of action, using some form of closed-ended response. Sociologist Harry Triandis and colleagues developed a series of scenarios (vignettes) designed to measure respondents' cultural orientation. One vignette simply stated, *You and your friends decided spontaneously to go out to dinner at a restaurant. What do you think is the best way to handle the bill?* Four response options were provided: (a) *split it equally, without regard to who ordered what*, (b) *split it according to how much each person makes*, (c) *the group leader pays the bill or decides how to split it*, (d) *compute each person's charge according to what that person ordered.* For this particular measure, respondents were asked to rank their top two options. This particular study consisted of a series of 16 different vignettes, each with a unique set of response options, all designed to assess some aspect of cultural orientation.

Vignette questions are flexible and can be used to measure many different types of attitudes and beliefs or scenarios. They are especially useful when trying to measure complex concepts that may be best described by way of example.

Attitude measures may be classified in a variety of different ways, but one classic taxonomy remains remarkably useful. In the 1960s, Stuart Cook and Claire Selltiz classified measures in the following taxa: self-report, physiological, behavioral, partially structured, and measures based on the performance of objective tasks. Self-report, physiological, and behavioral measures remain conceptually and operationally (except for technological advances in the case of physiological measures) similar to those used nearly half a century ago. Measures based on the performance of objective tasks are those in which the respondent is asked to perform some task that is expected to be influenced by the respondent's attitude (e.g., when asked to draw a picture of a quarter, a respondent who has more favorable attitudes toward money may be expected to draw the quarter larger).

Partially structured measures are those in which respondents are presented with an ambiguous stimulus and asked to respond as they see fit. Responses to partially structured (e.g., vignette) measures are influenced by pertinent characteristics (e.g., attitudes) of the respondents and can therefore be used to infer characteristics about the respondents. One advantage of vignette measures is that they are more indirect than self-report measures. Although the topic under consideration is clearly present in the vignettes, responses to partially structured vignette questions do not require the intentional recollection of stored information, as with direct questions. Thus, vignette questions can be used to assess characteristics of the respondents that they may be either unwilling or unable to admit, even to themselves.

An example of this use may be seen in the psychology literature with an "older cousin" of the vignette question. This is the Thematic Apperception Test, in which respondents are presented with an actual illustration or pictographic image and are asked to write an imaginative story about the graphic. Respondents' stories are coded by trained judges and are used to infer motives, such as the needs for achievement, affiliation, or power. Respondents' achievement, affiliation, or power motives are generally considered to be consciously inaccessible to respondents and thus only available to researchers via administration of the Thematic Apperception Test.

Partially structured vignette measures may also be used with closed-ended response options. Patrick Vargas and his colleagues presented respondents with a series of vignettes describing different protagonists engaging in ambiguously conflicted behaviors; for example, in assessing attitudes toward being religious, one vignette described a woman who declared herself to be very religious but also admitted to not having attended religious services since starting college. Respondents were asked to judge how religious they believed the protagonists to be using 11-point scales. Respondents tended to contrast their judgments of the protagonists away from their own beliefs: Very religious people judged the woman to be more atheistic, whereas nonreligious people judged the woman to be more religious. Similar results were found with different vignettes assessing attitudes toward dishonesty, inequality, and politics. Of note, in all studies, responses to the vignette questions were uncorrelated with direct measures of the construct under consideration, and responses to the vignette questions reliably predicted both self-reported and actual behaviors, beyond what could be predicted by self-report measures.

The Internet, as a mode of survey data collection, is especially conducive for the deployment of vignette questions, as the vignettes can be displayed to respondents as text, with visuals, with audio stimuli, or all of these aids.

Patrick Vargas

See also Attitude Measurement; Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Experimental Design; Factorial Design; Factorial Survey Method (Rossi's Method); Independent Variable; Interaction Effect; Main Effect; Random Assignment; Web Survey

Further Readings

Cook, S. W., & Selltiz, C. (1964). A multiple-indicator approach to attitude measurement. *Psychological Bulletin*, *62*, 36–55.

Triandis, H. C., Chen, X. P., & Chan, D. K.-S. (1998). Scenarios for the measurement of collectivism and individualism. *Journal of Cross-Cultural Psychology*, *29*, 275–289.

Vargas, P. T., von Hippel, W., & Petty, R. E. (2004). Using "partially structured" attitude measures to enhance the attitude-behavior relationship. *Personality and Social Psychology Bulletin*, *30*, 197–211.

Visual Communication

Visual communication involves the transmission of information through the visual sensory system, or the eyes and sense of sight. In surveys, visual communication relies heavily on verbal communication (i.e., written text) but can also include nonverbal communication (i.e., through images conveying body language, gestures, or facial expressions) and paralinguistic communication (i.e., through graphical language). Visual communication can be used to transmit information independently or in combination with aural communication. When conducting surveys, the mode of data collection determines whether information can be transmitted visually, aurally, or both. Whether survey information is transmitted visually or aurally influences how respondents first perceive and then cognitively process information to provide their responses.

Visual communication can consist of not only verbal communication but also nonverbal and paralinguistic forms of communication, which convey additional information that reinforces or modifies the meaning of written text. Nonverbal communication (i.e., information transferred through body language, gestures, eye contact, and facial expressions) is most common in face-to-face surveys but can also be conveyed with graphic images in paper and Web surveys. *Paralinguistic communication* is traditionally thought of as information transmitted aurally through the speaker's voice (e.g., quality, tone, pitch, inflection). However, recent literature suggests that graphical features, such as layout, spacing, font size, typeface, color, and symbols, that accompany written verbal communication can serve the same functions as aural paralinguistic communication. That is, these graphical features and images, if perceived and cognitively processed, can enhance or modify the meaning of written text in paper and Web surveys.

The process of perceiving visual information can be divided into two broad and overlapping stages: pre-attentive and attentive processing. In *pre-attentive processing*, the eyes quickly and somewhat effortlessly scan the entire visual field and process abstract visual features such as form, color, motion, and spatial position. The eyes are then drawn to certain basic visual elements that the viewer distinguishes as objects from other competing elements in the visual field that come to be perceived as background. Once the figure/ground composition is determined, the viewer can start discerning basic patterns among the objects. To distinguish such patterns, the viewer uses the graphical elements of proximity, similarity, continuity and connectedness, symmetry, and closure—as described by Gestalt psychology. During pre-attentive processing, the viewer uses a bottom-up model of processing where only stimuli from the visual field itself influence how objects and images are perceived. It is at this pre-attentive stage that graphical elements, strategically provided by survey designers, can help respondents perceive the basic navigational flow of the survey, including start of the survey, question order, and question groupings.

As respondents begin to recognize visual elements and patterns, *attentive processing* begins where they focus on the task of answering each question. Attentive processing is task oriented and involves the narrowing of one's vision to a limited visual field of attention known as the *foveal view*, which is 8 to 10 characters in width. During this stage, information is processed more slowly and often in a sequential order. It is also usually processed more deeply and committed to long-term memory rather than just reaching working memory as in pre-attentive processing. During attentive processing the viewer uses a top-down model of processing where his or her prior knowledge, experiences, and expectations about the particular situation influences how objects and images are perceived. It is at this stage that additional instructions or visual elements put in place to help facilitate the task of attending to, and answering, each individual question become useful to respondents. To be most effective, these elements should be located in the foveal view of attention near where they will be used.

Because visual and aural communication differ in how information is presented to survey respondents, the type of communication impacts how respondents initially perceive survey information. This initial step of perception influences how respondents cognitively process the survey in the remaining four steps (comprehension, retrieval, judgment formation, and answer reporting). Whereas paper surveys rely solely on visual communication, both Web and face-to-face surveys can utilize visual and aural communication. Web surveys rely extensively on visual communication and also have the potential to incorporate computer-mediated aural communication. In contrast, face-to-face surveys rely mostly on aural communication with the more occasional use of visual communication through show cards or other visual aids. Finally, telephone surveys rely on aural communication and generally do not incorporate any visual communication.

The specific effects that visual communication has on perception and the cognitive processing of information can contribute to mode effects during the survey response process. For example, visual communication allows the respondent to process a greater amount of information at once than does aural transmission; thus, survey designers can use more complex questions and response scales. However, the same questions and scales are not appropriate in aurally based surveys, and as a result, responses to survey modes may differ significantly.

Leah Melani Christian
and Jolene D. Smyth

See also Aural Communication; Comprehension; Gestalt Psychology; Graphical Language; Mode Effects; Mode of Data Collection; Retrieval

Further Readings

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed., 2007 update). Hoboken, NJ: Wiley.

Jenkins, C., & Dillman, D. A. (1997). Towards a theory of self-administered questionnaire design. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, et al. (Eds.), *Survey measurement and process quality* (pp. 165–196). New York: Wiley-Interscience.

Ware, C. (2004). *Information visualization: Perception for design*. San Francisco: Morgan Kaufmann.

Voice over Internet Protocol (VoIP) and the Virtual Computer-Assisted Telephone Interview (CATI) Facility

Voice over Internet Protocol (VoIP) refers to real-time telecommunication using the Internet to carry the digitized versions of the voices that are speaking to each other. The world of telecommunications has been turned on its ear in the past few years with VoIP telephony challenging the regional telephone companies as broadband connections provided by cable TV vendors penetrate the market. This technological change has the potential to be a force to redesign computer-assisted telephone interviewing (CATI) and multi-modal survey interviewing efforts. For example, starting in 2002, the Center for Human Resource Research (CHRR) at Ohio State University began implementing such a redesign.

A number of survey organizations have converted their office phone systems to VoIP using large, well-established vendors of proprietary systems. These systems tend to be more costly to deploy than those obtainable from smaller companies offering a more open architecture. Moreover, because the "big name" systems are proprietary and targeted at the office environment, they are often harder to adapt and customize for survey research applications. A more open architecture allows for more creativity in designing the integration of survey interviewing and telephony.

Modern computing emphasizes the use of the Internet and relational database techniques. The CAPI (computer-assisted personal interviewing) and CATI technology used at CHRR for the past 2 years utilizes these technologies. For CATI, the client on the interviewer's machine communicates with the server at CHRR, and for CAPI, the client and "server" (an application that runs on several laptops) are both present on the interviewer's computer. This makes CATI/CAPI not two software systems but one. Whether for financial transaction processing, online shopping, or CATI interviewing, relational database tools are designed to support a high volume of queries and transaction processing. If a CATI interviewer can access a remote server via the Web from the telephone research facility, they can do so from home or from thousands of miles away. If they have a broadband connection, they can readily and inexpensively be connected to a VoIP system with a "softphone" (software that handles all the steps in placing or answering a call, and a USB connection for the interviewer's headset). The Internet carries the conversation from the interviewer's location to the central office where it crosses over to the public switched telephone network. A Web-based dialer schedules calls, presents cases to the interviewer, and places the calls.

With a VoIP connection to an interviewer's home (or to a research call center), he or she can log, record, and monitor calls just as in other traditional call centers. By using a client-server approach, the data are stored only in the central office, never at the remote location. Interviewers can be trained over the Internet, with utilities that allow the trainees to view the instructor's monitor as he or she instructs them on the use of the system and on proper interviewing techniques. During the venerable round-robin parts of interviewer training, where each trainee takes turns acting as the interviewer and the instructor acts as the respondent, the same utility allows the instructor to view the monitor of each trainee. As soon as someone falls behind or fails to follow directions, the trainer notices the nonconforming screen and brings that trainee back with the rest of the group. A single trainer can handle the groups of approximately eight trainees quite readily.

The VoIP branch of the call connecting the interviewer and the respondent is more secure than the connection from the central office to the respondent that is used over a public switched telephone network landline. First of all, VoIP converts the audio signal into data packets using a *codec*—a mathematical rule for digitizing and compressing audio signals. Unless one knows the codec—and there are many—one cannot reconvert the digital packets back to sound. Moreover, there are many ways of encrypting the digital packets, forcing a would-be eavesdropper to both break the encryption and reverse-engineer the codec. Second, the VoIP system conceals and encrypts the information on who is being called. The dialer communicates with the rest of the system over a separate data stream, which is itself encrypted, and the packets of data that make the connection between the interviewer and the central office VoIP server are channeled through separate servers using secure tunneling protocols. Using separate servers and data streams to handle the signaling that sets up the call from the data stream carrying the packets further complicates the task of a would-be eavesdropper. Cracking the voice packets does not allow one to know who is talking. There are rumors that some VoIP systems are so arcane and their functioning so ill-understood by outsiders that calls are, for practical purposes, immune from being logged let alone recorded by outsiders. Contrast these complications associated with compromising a VoIP conversation with what it takes to intercept a conversation over the public switched telephone network—calls that are well known to be vulnerable to intercept.

CHRR has used this virtual CATI shop technology for a variety of surveys including the National Longitudinal Surveys. The technology has several advantages. First, the virtual CATI facility technology allows survey organizations to use more effectively those remote interviewers who work in sparsely populated sampling areas. When not working face-to-face cases, they can work in a virtual CATI facility and achieve the same productivity metrics as in that more controlled environment. This technology will allow more survey organizations to train and use a skilled national field force because the volume of telephone work will keep remote staff employed and on the payroll between face-to-face engagements.

Second, this technology is valuable for tracking interviewer behavior for face-to-face surveys when the interviewers use the phone for contacting and making appointments. The technology allows for the tracking of interviewer calls to determine whether they are working as efficiently as they should. However, some interviewers shirk, and shirking can come in the form of not working when they say they are working or not placing the calls they say they are placing. When interviewers call from home over landlines, they currently cannot be monitored. With VoIP, they can be monitored and measured. Better monitoring will improve both technique and productivity.

Third, using VoIP allows for the implementation of sophisticated audio computer-assisted self-interviewing techniques by staff working from home on telephone interviews. The flexibility of VoIP and an open architecture allow the interviewer to switch the respondent to a system that uses voice recordings to ask a question and then, using voice recognition methods, record the answer and do the necessary branching.

Finally, this technology reduces the need to maintain a costly "bricks and mortar" call center to house all the call agents while still retaining the traditional call center's advantages in terms of monitoring, scheduling, and training. Although some call agents may not have a home environment that can serve as a virtual call center, many will. The virtual call center technology expands the labor pool (including the aged and infirm) and allows for the more flexible deployment of professional interviewers.

Randall Olsen and Carol Sheets

See also Audio Computer-Assisted Self-Interviewing (ACASI); Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Self-Interviewing (CASI); Computer-Assisted Telephone Interviewing (CATI); Data Management; Interviewer Monitoring

Further Readings

Bocklund, L., & Bengtson, D. (2002). *Call center technology demystified: The no-nonsense guide to bridging customer contact technology, operations and strategy.* Annapolis, MD: Call Center Press.

Olsen, R. J. (2004, May). *Does the Web doom traditional call centers?* Poster session presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Voluntary Participation

Voluntary participation refers to a human research subject's exercise of free will in deciding whether to participate in a research activity. International law, national law, and the codes of conduct of scientific communities protect this right. In deciding whether participation is voluntary, special attention must be paid to the likely participants' socioeconomic circumstances in determining which steps must be put in place to protect the exercise of free will. The level of effort involved in clarifying voluntariness is not fixed and depends on several circumstances, such as the respondents' abilities to resist pressures like financial inducements, authority figures, or other forms of persuasion. Special care, therefore, must be taken to eliminate undue pressure (real and perceived) when research subjects have a diminished capacity to refuse.

Basic Requirements

The essential qualities of "voluntariness" include the following:

- The subject has a choice to participate.
- The choice is made without coercion.
- The choice is made without undue influence.
- The subject has foreknowledge of risks or benefits.

For a choice to occur, subjects must be of age and have the mental capacity to make such decisions. When this assumption cannot be upheld, this decision is left to the subjects' parents or legal guardians. For general-interest telephone surveys this prerogative may be largely implied, whereas in other situations it may have to be stated very precisely. The research director has the obligation to provide this choice at the outset, and the subject enjoys the prerogative to withdraw after data collection has begun.

Coercion refers to a threat that some type of harm will come to the research subject unless he or she participates. Such coercion might include an overt threat of violence, injury to reputation, or implied damage to career status. It may be important in some settings to take special precautions to ensure that participants do not perceive that they will face some retaliation from authority figures if they decline to participate.

Undue influence refers to excessive, improper, or immoral benefits. A frequent point of attention is cash incentives. To avoid undue influence, an incentive must be nominal, or so small that the subject can afford to refuse. This consideration is especially important when the indigent or children are the focus of data collection. The concept of voluntariness encompasses the realization that participants may perceive that persons in authority will reward their participation. Therefore, in some situations it may be necessary to take steps to dispel this explicitly.

For participation to be voluntary, subjects must have foreknowledge of likely risks and benefits of participation and of their option to withdraw from participation at any time. Subjects should have the opportunity to consider how the collected data will be used, whether confidentiality is being protected, who is sponsoring the research, how long they are expected to participate, and what costs or benefits they may expect from participation, refusal, or withdrawal. This means that information provided to subjects is without deceit about the project.

Some Exceptions

These requirements apply to research and should not be confused with taking tests, completing forms for employment, or filling out applications for public benefits.

Some data-collection activities are compulsory. Among these is the decennial Census of the United States and the Annual Survey of U.S. Direct Investment Abroad, the South Korean Monthly Industrial Production Survey, the British Surveillance of Healthcare Associated Infections, or the Canadian Survey of Innovations. The exceptions to voluntariness are made explicit in the respective countries' legislation.

Sean O. Hogan

See also Confidentiality; Deception; Ethical Principles; Informed Consent; Institutional Review Board (IRB)

Further Readings

American Association for Public Opinion Research. (2005). *Code of professional ethics and practices*. Retrieved April 24, 2008, from http://www.aapor.org/aaporcodeofethics

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Retrieved January 2, 2007, from http://www.apa.org/ethics/code2002.html#8

The Nuremberg Code: Trials of war criminals before the Nuremberg military tribunals under control council Law No. 10. (1949–1953). Washington, DC: Government Printing Office. Retrieved January 2, 2007, from http://www.ushmm.org/research/doctors/Nuremberg_Code.htm

U.S. Department of Health and Human Services. Protection of Human Subjects (45 C.F.R. 46). Retrieved January 2, 2007, from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

U.S. Department of Health, Education, and Welfare. (1979, April 18). *The Belmont report*. Retrieved January 2, 2007, from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm

World Medical Association. (2004). *Helsinki Declaration: Ethical principles for medical research involving human subjects*. Retrieved January 2, 2007, from http://www.wma.net/e/policy/b3.htm

W

WAVE

Within the context of survey research, a wave refers to each separate survey in a series of related surveys. If a survey is conducted only once, then the concept of a "wave" does not apply. It is when a survey is conducted two or more times, for example, once a year for 5 years in a row, that each repeated survey is called a *wave*. These waves may represent a panel survey in which the same respondents are tracked over time, often being asked some or all of the same questions in each wave. Other multi-wave (longitudinal) surveys use independent samples at each wave.

When a panel study is conducted and respondents who are interviewed in a prior wave are not successfully interviewed in a subsequent wave, even though they should have been, then *panel attrition* results. Common reasons for panel attrition at subsequent survey waves are refusals, moving, and illness or death. Some long-term survey designs rotate respondents in and out of being interviewed at various waves. Other long-term designs eliminate a random portion of the respondents that had been interviewed in previous waves and replace them with a random sample of "first time" respondents who likely will be contacted again in future waves.

Paul J. Lavrakas

See also Attrition; Longitudinal Studies; Panel; Panel Survey

Further Readings

Kasprzyk, D., Duncan, G., Kalton, G., & Singh, M. P. (1989). *Panel surveys.* New York: Wiley.

WEB SURVEY

Web surveys allow respondents to complete questionnaires that are delivered to them and administered over the World Wide Web. This offers advantages over other survey methods but also poses problems for quality data collection.

A main advantage of Web surveys, compared with other survey methods, is lower costs. Lower costs enable researchers to collect larger samples than affordable with other survey methods and have made Web surveys an option for many who cannot afford other survey methods. Technology allowing for complex questionnaire designs using features available in computer-assisted interviewing programs (such as CAPI, CASI, and CATI) is inexpensive. There are other advantages to Web surveys. The ability to quickly contact respondents over the Internet and process questionnaires through Web sites can shorten field periods compared to other survey methods. Web surveys can be used in conjunction with other methods, giving respondents another option for survey participation. The Web's role as an electronic meeting place has increased the ability to survey some rare, isolated, or marginal populations. In principle it is possible to conduct completely anonymous Web surveys, which may make surveying some populations easier.

Web surveys, however, have limitations that prevent them from being a simple substitute for other survey methods. A major limitation of Web surveys is their inability to represent the general population, or even the population of Internet users, without special efforts or techniques. Internet access, a requirement for participation in Web surveys, is more common among those with higher, as opposed to lower, socioeconomic status. This bias means that Web surveys are likely to have substantial coverage error for many survey topics (e.g., those related to health, finances, and education) if they are intended to represent the general population. Further, sampling those who have Internet access is problematic. There is no single sampling frame of the population of Internet users, and it is not possible to construct one because there is no standard system for assigning email addresses, the Internet counterpart to street addresses or telephone numbers.

Researchers have different approaches to dealing with these coverage and sampling issues. Mick Couper has categorized Web surveys according to their methods of respondent selection. Some approaches do not use probability sampling techniques for initial selection of respondents. A common approach is to rely on high traffic through Web sites to produce convenience samples. Another approach is to use Web sites to recruit pools of potential participants who are later sampled for Web surveys. Demographic information about those in a pool can be used for sampling quotas or for post-survey data weighting. These approaches rely on potential participants to opt into a survey or pool, which can produce a self-selection bias. The ability of such samples to represent a full population of Internet users is controversial. Other Web survey approaches use probability sampling methods for respondent selection. When Internet use is high among a population, email lists can serve as sampling frames (e.g., all students, staff, and faculty at a university). Web surveys based on these list samples are potentially representative of those populations. Another approach is to systematically sample users of Web sites using intercept or "pop-up" surveys; those data are potentially representative of users of those Web sites. A probability sampling approach for the population of Internet users could use random-digit dialing telephone surveys or in-person surveys of the general population to recruit Internet users who will then be sampled for Web surveys. Those without Internet access are considered out of the sample. Web surveys based on such samples are potentially representative of the population of Internet users. Finally, the last approach can be modified for conducting Web surveys of the general population. For example, random-digit dialing telephone surveys of a population are used to recruit potential participants for Web surveys. Those without Internet access are not considered out of the sample but are instead provided with Internet access and technology in order to participate in Web surveys. Web survey data collected from such samples are potentially representative of a population.

Web surveys may have unique sources of survey error that researchers must consider when evaluating data. For example, unique measurement error may be associated with the way Web questionnaires are displayed or controlled. Differences in Web browsers and display settings may alter the layout of questions, which could affect how respondents process them. Failing to answer questions or backing up through a questionnaire may disrupt programmed skip patterns, leading respondents to answer questions not meant for them or to skip questions they should receive. Web surveys may have unique sources of nonresponse error. Potential respondents may not be able to participate in a survey because their computer hardware or software may be inadequate for, or incompatible with, some Web survey programs. Web surveys that require high levels of computer literacy may discourage low-literacy respondents. Email requests for survey participation may be rejected or ignored as spam, preventing participation.

Finally, Web surveys may pose additional participant confidentiality issues. Security breaches of computer data, Web sites, and Internet communications are common and may threaten the confidentiality of Web survey participants. Measures should be in place to maximize respondent confidentiality.

Lew Horner

See also Computer-Assisted Self-Interviewing (CASI); Convenience Sampling; Coverage Error; Measurement Error; Nonresponse Error; Sampling Frame; Self-Selection Bias

Further Readings

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

Dillman, D. A., & Bowker, D. K. (2001). The Web questionnaire challenge to survey methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 159–178). Lengerich, Germany: Pabst Science.

WEIGHTING

Weighting is a correction technique that is used by survey researchers. It refers to statistical adjustments that are made to survey data after they have been collected in order to improve the accuracy of the survey estimates. There are two basic reasons that survey researchers weight their data. One is to correct for unequal probabilities of selection that often have occurred during sampling. The other is to try to help compensate for survey nonresponse. This entry addresses weighting as it relates to the second of these purposes.

Why Weighting for Nonresponse Is Necessary

Essentially all surveys suffer from nonresponse. This occurs either when elements (persons, households, companies) in the selected sample do not provide the requested information or when the provided information is useless. The situation in which all requested information on a sampled element is missing is called *unit nonresponse.*

As a result of unit nonresponse, estimates of population characteristics may be biased. This bias occurs if (a) some groups in the population are over- or underrepresented in the sample because of their differential response/nonresponse rates and (b) these groups are different with respect to the variables being measured by the survey. As a consequence, wrong conclusions are drawn from the survey results.

It is vital to try to reduce the amount of nonresponse in the field work as much as possible. Nevertheless, in spite of all these efforts, a substantial amount of nonresponse usually remains. To avoid biased estimates, some kind of correction procedure must be carried out. One of the most important correction techniques for nonresponse is weighting. It means that every observed object in the survey is assigned a weight, and estimates of population characteristics are obtained by processing weighted observations instead of the (unweighted) observations themselves.

Basics of Weighting to Correct for Nonresponse

Suppose that the objective of a survey is to estimate the population mean, $\bar{Y}$, of a variable Y. Suppose further that a simple random sample of size n is selected with equal probabilities and without replacement from a population of size N. The sample can be represented by a series of N indicators $t_1, t_2, \ldots, t_N$, where the k-th indicator t_k assumes the value 1 if element k is selected in the sample, and otherwise it assumes the value 0. In case of complete response, the sample mean,

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} t_k Y_k, \quad (1)$$

is an unbiased estimator of the population mean. In case of nonresponse, this estimator may be biased. Assuming that every element, k, in the population has an unknown probability, ρ_k, of response when invited to participate in the survey, the bias of the mean $\bar{y}_R$ of the available observations is equal to

$$B(\bar{y}_R) = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}, \quad (2)$$

where $R_{\rho Y}$ is the correlation between the values of the survey variable and the response probabilities, S_Y is the standard deviation of the Y, and S_ρ is the standard deviation of the response probabilities.

Weighting is a frequently used approach to reduce this nonresponse bias. Each observed element, k, is assigned a weight, w_k. Thus, the response mean, $\bar{y}_R$, is replaced by a new estimator,

$$\bar{y}_W = \frac{1}{n}\sum_{k=1}^{N} w_k t_k Y_k. \quad (3)$$

Correction weights are the result of the application of some weighting technique. The characteristics of the correction weights should be such that the weighted estimator has better properties than the unweighted response mean.

Weighting is based on the use of auxiliary information. *Auxiliary information* is defined as a set of variables that have been measured in the survey and for which information on the population distribution (or the complete sample) is available. By comparing the population parameters for an auxiliary variable with its observed response distribution in the survey, it can be determined whether or not the sample is representative of the population (with respect to this variable). If these distributions differ in non-negligible ways, one must conclude that nonresponse has resulted in a biased sample.

The auxiliary information is used to compute adjustment weights. Weights are assigned to all records of the observed elements. These weights are defined in such a way that population characteristics for the auxiliary variables can be computed without error. So when weights are applied to the estimate population means of the auxiliary variables, the estimates must be equal to the true values, that is,

$$\bar{x}_W = \frac{1}{N}\sum_{k=1}^{N} w_k t_k X_k = \bar{X}. \qquad (4)$$

If this condition is satisfied, then the weighted sample is said to be *representative* with respect to the auxiliary variable used.

If it is possible to make the sample representative with respect to several auxiliary variables, and if these variables have a strong relationship with the survey variables, then the (weighted) sample will also be (approximately) representative with respect to these variables, and hence estimates of population characteristics will be more accurate.

Post-Stratification

The most frequently used weighting technique to correct for the effects of nonresponse is post-stratification. One or more qualitative auxiliary variables are needed to apply this technique. In the following explanation, only one such variable is considered. Extension to more variables is straightforward.

Suppose there is an auxiliary variable, X, having L categories. X divides the population into L strata. The number of population elements in stratum h is denoted by N_h, for $h = 1, 2, \ldots, L$. Hence, the population size is equal to $N = N_1 + N_2 + \cdots + N_L$.

Post-stratification assigns identical correction weights to all elements in the same stratum. The weight w_k for an element k in stratum h (in the case of a simple random sample with equal probabilities) is defined by

$$w_k = \frac{N_h/N}{n_h/n}, \qquad (5)$$

where n_h is the number of respondents in stratum h. If the values of the weights are substituted in Equation (4), the result is the well-known post-stratification estimator,

$$\bar{y}_{ps} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h, \qquad (6)$$

where $\bar{y}_h$ is the response mean in stratum h. So, the post-stratification estimator is equal to a weighted sum of response means in the strata.

The nonresponse bias disappears if there is no relationship between response behavior and the survey variable within strata. Two situations can be distinguished in which this is the case:

- The strata are homogeneous with respect to the target variable; that is, this variable shows little variation within strata.
- The strata are homogeneous with respect to the response behavior; that is, response probabilities show little variation within strata.

Linear Weighting

Even in the case of full response, the precision estimators can be improved if suitable auxiliary information is available. Suppose there are p auxiliary variables. The values of these variables for element k are denoted by the vector $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$. The vector of population means is denoted by $\bar{X}$. If auxiliary variables are correlated with the survey variable, Y, then for a suitably chosen vector $B = (B_1, B_2, \ldots, B_p)'$ of regression coefficients for a best fit of Y on X, the residuals, $E_k = Y_k - X_k B$, vary less than the values of target variable itself. The ordinary least squares solution B can be estimated, in the case of full response, by

$$b = \left(\sum_{k=1}^{N} t_k X_k X_k'\right)^{-1}\left(\sum_{k=1}^{N} t_k X_k Y_k\right). \qquad (7)$$

The generalized regression estimator is now defined by

$$\bar{y}_{GR} = \bar{y} + (\bar{X} - \bar{x})' b, \qquad (8)$$

in which $\bar{x}$ is the vector of sample means of the auxiliary variables. The generalized regression estimator is *asymptotically design unbiased.* This estimator reduces the bias caused by nonresponse if the underlying regression model fits the data well.

The generalized regression estimator can be rewritten in the form of the weighted estimator (4), where the correction weight, w_k, for observed element k is equal to $w_k = v'X_k$, and v is a vector of weight coefficients, which is equal to

$$v = \left(\sum_{k=1}^{N} \frac{t_k X_k X_k'}{n} \right)^{-1} \bar{X}. \quad (9)$$

Post-stratification turns out to be a special case of linear weighting. If the stratification is represented by a set of dummy variables, where each dummy variable denotes a specific stratum, Equation 8 reduces to Equation 6.

Linear weighting can be applied in more situations than post-stratification. For example, post-stratification by age, class, and sex requires that the population distribution for the crossing of age and class by sex be known. If only the separate marginal population distributions of age, sex, or both, are known, then post-stratification cannot be applied. In that case in this example, only one variable can be used. However, linear weighting makes it possible to specify a regression model that contains both marginal distributions. In this way more information is used, and this generally will lead to better survey estimates.

Linear weighting has the disadvantage that some correction weights may turn out to be negative. Such weights are not wrong but simply a consequence of the underlying theory. Usually, negative weights indicate that the regression model does not fit the data well. Some analysis packages are able to work with weights, but they do not accept negative weights. This may be a reason not to apply linear weighting.

Multiplicative Weighting

Correction weights produced by linear weighting are the sum of a number of weighted coefficients. It is also possible to compute correction weights in a different way, namely, as the product of a number of weight factors. This weighting technique is usually called *raking* or *iterative proportional fitting*. Here this process is denoted by *multiplicative weighting*, because weights are obtained as the product of a number of factors contributed by the various auxiliary variables.

Multiplicative weighting can be applied in the same situations as linear weighting as long as only qualitative variables are used. Correction weights are the result of an iterative procedure. They are the product of factors contributed by all cross-classifications. To compute weight factors, the following scheme has to be carried out:

1. Introduce a weight factor for each stratum in each cross-classification term. Set the initial values of all factors to 1.
2. Adjust the weight factors for the first cross-classification term so that the weighted sample becomes representative with respect to the auxiliary variables included in this cross-classification.
3. Adjust the weight factors for the next cross-classification term so that the weighted sample is representative for the variables involved. Generally, this will disturb the representativeness of the other cross-classification terms in the model.
4. Repeat this adjustment process until all cross-classification terms have been dealt with.
5. Repeat steps 2, 3, and 4 until the weight factors do not change by any more than a negligible amount.

Multiplicative weighting has the advantage that computed weights are always positive. It has the disadvantage that there is no clear model underlying the approach. Moreover, there is no simple and straightforward way to compute standard errors of the weighted estimates. Linear weighting is based on a regression model, which allows for computing standard errors.

Calibration Estimation

Jean-Claude Deville and Carl-Erik Särndal have created a general framework for weighting of which linear weighting and multiplicative weighting are special cases. This framework is called *calibration.* Assuming simple random sampling with equal probabilities, the starting point is that adjustment weights have to satisfy two conditions:

1. The adjustment weights w_k have to be as close as possible to 1.
2. The weighted sample distribution of the auxiliary variables has to match the population distribution; that is,

$$\bar{x}_W = \frac{1}{n}\sum_{k=1}^{N} t_k w_k X_k = \bar{X}. \tag{10}$$

The first condition sees to it that resulting estimators are unbiased, or almost unbiased, and the second condition guarantees that the weighted sample is representative with respect to the auxiliary variables used.

A distance measure, $D(w_k, 1)$, that measures the difference between w_k and 1 in some way, is introduced. The problem is now to minimize

$$\sum_{k=1}^{N} t_k D(w_k, 1) \tag{11}$$

under the condition (10). This problem can be solved by using the method of Joseph Lagrange. By choosing the proper distance function, both linear and multiplicative weighting can be obtained as special cases of this general approach. For linear weighting the distance function D is defined by $D(w_k, 1) = (w_k - 1)^2$, and for multiplicative weighting the distance $D(w_k, 1) = w_k \log(w_k) - w_k + 1$ must be used.

Other Issues

There are several reasons why a survey statistician may want to have control over the values of the adjustment weights. One reason is that extremely large weights are generally considered undesirable. Use of such weights may lead to unstable estimates of population parameters. Another reason to have some control over the values of the adjustment weights is that application of linear weighting might produce negative weights.

The calibration approach allows for a weighting technique that keeps the adjustment weights within pre-specified boundaries and, at the same time, enables valid inference. Many surveys have complex sample designs. One example of such a complex design is a cluster sampling. Many household surveys are based on cluster samples. First, a sample of households is selected. Next, several or all persons in the selected households are interviewed. The collected information can be used to make estimates for two populations: the population consisting of all households and the population consisting of all individual persons. In both situations, weighting can be carried out to correct for nonresponse. This results in two weights assigned to each record: one for the household and one for the individual. Having two weights in each record complicates further analysis.

The generalized regression estimation offers a solution. The trick is to sum the dummy variables corresponding to the qualitative auxiliary variables for the individuals over the household. Thus, quantitative auxiliary variables are created at the household level. The resulting weights are assigned to the households. Furthermore, all elements within a household are assigned the same weight, and this weight is equal to the household weight. This approach forces estimates computed using the element weights to be consistent with estimates based on the cluster weights.

Jelke Bethlehem

See also Auxiliary Variable; Bias; Cluster Sample; Differential Nonresponse; Nonresponse; Post-Stratification; Probability of Selection; Raking; Representative Sample; Unit Nonresponse

Further Readings

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Bethlehem, J. G. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 275–288). New York: Wiley.

Bethlehem, J. G., & Keller, W. J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, *3*(2), 141–154.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*, 1013–1020.

WesVar

WesVar is a Windows-based software package for analyzing data collected by surveys with complex sample designs. Sample designs that can be analyzed include stratified or unstratified designs, single- or multi-stage designs, one-time or longitudinal designs. A key aspect of WesVar is the calculation of design-based standard errors that take the complex sample design (e.g., stratification and clustering) into account.

WesVar calculates sampling variances using the jackknife and balanced repeated replication methods. WesVar users can calculate and use their own replicate weights or have WesVar calculate the replicate weights. WesVar can perform weight adjustments for nonresponse, post-stratification, and raking. Such weight adjustments are performed on both the full-sample weights and the replicate weights, so that calculated sampling variances take into account the effects of weight adjustments.

The estimates and standard errors calculated by WesVar take into account the sample design and adjustments to sampling weights. Statistical software that ignores the sample design and assumes simple random sampling typically underestimates standard errors.

WesVar can calculate all of the following:

- Multi-way tables containing totals, means, or proportions, along with estimated standard error, coefficient of variations, confidence intervals, and design effects
- Estimates of medians and other quantiles, along with estimated standard errors
- Complex functions of estimates, such as ratios, differences of ratios, and log-odds ratios, along with estimated standard errors
- Chi-square tests of independence for two-way tables of weighted frequencies
- Estimated coefficients and their standard errors for linear and logistic regression models and associated significance tests for linear combinations of parameters

WesVar was developed and is distributed by Westat. The program and user manual can be downloaded from the WesVar Web site.

Richard Sigman

See also Balanced Repeated Replication (BRR); Complex Sample Surveys; Design Effects (*deff*); Jackknife Variance Estimation; Multi-Stage Sample; Replicate Methods for Variance Estimation; Variance Estimation

Further Readings

Westat. (2007). *WesVar 4.3 user's guide*. Retrieved September 13, 2007, from http://www.westat.com/westat/wesvar/about/WV_4-3_Manual.pdf

WesVar: http://www.westat.com/westat/wesvar/index.html

WITHIN-UNIT COVERAGE

Within-unit coverage refers to the accuracy of selection of the respondent to be interviewed after a household is contacted, generally by phone or face-to-face methods, and the informant in the household identifies the prospective respondent, according to the interviewer's instructions. Theoretically, each eligible person living in the unit (e.g., all adults) should have a known and nonzero chance of selection. A number of probability, quasi-probability, and nonprobability methods exist for choosing the respondent, and only the probability methods allow estimation of chances of selection. Humans, however, do not always behave according to the logic of probability methods, and this leads to errors of within-unit coverage. Missed persons within households contribute more to noncoverage of the Current Population Survey, for example, than do missed housing units.

In general, methods that do not nominate the correct respondent by sex and age are more likely to produce within-unit coverage errors than are those that do specify those characteristics. Those methods that ask for a full listing of the household's adults, such as the Kish technique, can be time consuming and threatening to some informants. The last-birthday and next-birthday procedures were devised to be nonintrusive and yield within-unit probability samples; yet, between about 10% and 25% of the time, the informant chooses the wrong respondent. This can happen because of misunderstanding the question, not knowing all the birthdays of household members, or the informant's deliberate self-selection instead of the designated respondent. If birthdays do not correlate with the topics of survey questions, this is not very problematic. Training interviewers rigorously and having them verify the accuracy of the respondent's selection can help to overcome difficulties with the birthday methods. Rigorous training also helps considerably with the Kish procedure, an almost pure probability method.

Another problem is undercoverage of certain kinds of people, especially minorities, the poor, the less educated, inner-city dwellers, renters, young males (particularly young black or Hispanic males), and older women in single-person households, as well as omission of some household members inadvertently or deliberately. For example, some informants conceal the presence of males in households that would cause the loss of welfare eligibility if wage-earning adults, particularly males, were known to live there. Other informants do not mention members who stay there most of the time but not all the time. In some populations the composition of households changes frequently; this happens with inner-city black and Hispanic communities, migrant workers, or migrant populations such as American Indians who can divide their time between cities and their reservation on a seasonal basis. Sometimes the informant and the survey planners have different definitions of the word *household*. Another source of error is informants' thinking the interviewer wants a count of families in the unit instead of all individuals in the unit. The larger the household is, the greater is the chance for noncoverage error (discrepancy between the true number of persons and the number obtained on the roster). Sometimes within-household coverage problems stem from counting individuals more than once.

In many situations, obtaining a roster of household members by initials or other ways allowing more anonymity can improve representation within households. Methods exist that do not require a list. In addition, making the target population clear to informants at the outset improves accuracy of selection.

Cecilie Gaziano

See also Kish Selection Method; Last-Birthday Selection; Within-Unit Selection

Further Readings

Brogan, D. J., Denniston, M. M., Liff, J. M., Flagg, E. W., Coates, R. J., & Brinton, L. A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, *153*, 1119–1127.

Lavrakas, P. J., Stasny, E. A., & Harpuder, B. (2000). A further investigation of the last-birthday respondent selection method and within-unit coverage error. *Proceedings of the Section on Survey Research Methods* (pp. 890–895). Alexandria, VA: American Statistical Association.

Maklan, D., & Waksberg, J. (1988). Within-household coverage in RDD surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 51–72). New York: Wiley.

O'Rourke, D., & Lakner, E. (1989). Gender bias: Analysis of factors causing male underrepresentation in surveys. *International Journal of Public Opinion Research*, *1*, 164–176.

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? Survey undercoverage and household roster questions. *Journal of Official Statistics*, *13*, 1–18.

WITHIN-UNIT COVERAGE ERROR

Within-unit coverage error refers to the bias, variance, or both, that may result when the respondents who are selected from within a sampling unit, and from whom data are gathered, differ in non-negligible ways from those who were missed from being selected but in theory should have been selected.

Many probability sample surveys are made up of a target population of individuals that belong to one sampling unit. But often these surveys use a two-stage design in which the unit is sampled first, and then a respondent (or more than one) from within the unit is selected to be interviewed. One example is a household or a unit within an organization, from which one person or a subset of persons per sampling unit is surveyed. These surveys usually select randomly one or more persons among all eligible persons that belong to a certain sampling unit according to some a priori definition. Within-unit coverage error can occur when one or more eligible persons within a selected sampling unit have a zero chance of selection, have a 100% (certain) chance of being selected, or have some other disproportionate chance of being selected, for example, because they belong to more than one sampling unit.

Within-unit coverage problems that might lead to error can therefore be defined as the difference between the sampling frame and the target population at the level of the individual residents. Within-unit undercoverage problems occur if people that are in fact eligible and should be counted as sampling unit members have no chance or less chance of being selected because they are not "recognized" by the informant or the selection technique as being a sampling unit member or they are not mentioned at all when sampling unit members are listed for the respondent selection process. The undercoverage problem arises because these persons are part of the target population but do not have a correct chance (probability) of being part of the frame population. Persons are said to be overcovered if they are listed in the frame population more than once and therefore have more than one chance to be selected into the sample but are only represented once in the target population.

Within-unit coverage error includes a bias as well as an error variance component. In both cases, within-unit coverage error is a property of a statistic, not a survey. Within-unit coverage bias in a statistic of interest occurs only if the members of the population that are not covered (or are overcovered) are different from those who are covered (or are only covered once) with regard to the statistic of interest. Therefore, there are two conditions that have to be met to produce within-unit coverage bias in a statistic of interest. The first condition is a difference between the sampling frame and the target population—subsets of the population are overcovered or undercovered. The second condition entails a difference of those people overcovered or undercovered from the rest of the target population with regard to the statistic of interest. Each condition is necessary but not sufficient to produce coverage error.

The literature on coverage error has described a variety of reasons for undercoverage as well as overcoverage problems. With regard to undercoverage, two major theories have emerged. The first theory is that the informant listing the members of the sampling unit usually does not know, or has difficulties understanding, the complicated definitions and rules, such as the de jure and de facto residence rules, that determine who should be counted as a member of the sampling unit. Therefore the informant might erroneously, and without any intent, not list people who, per the definition, belong to the sampling unit. To improve the clarity of these rules and definitions, interviewers can be asked to administer additional questions, after an initial household listing has been established, that focus on situations that usually lead to within-unit undercoverage problems. The second theory of undercoverage claims that informants might intentionally not list members of the sampling unit because they fear that something bad will happen if they do, for example, that illegal activities might be detected or that providing information about members of the sampling unit could have negative consequences with regard to monetary support by the government. People might just be uncomfortable and suspicious if they are asked to list sampling unit members by a survey organization that just made contact with them and not list them at all. Because listing sampling unit members of a household by name can be perceived as intrusive, survey organizations have allowed the informant to use initials when listing the sampling unit members. Other respondent selection methods that do not require listing household members, but simply select the respondent based on a criterion, have been found to be helpful in increasing the coverage of subgroups of the population usually undercovered when listing is used and also in reducing unit nonresponse that is more likely to result when overly intrusive within-unit selection methods are deployed. The last-birthday selection method is the most used of these respondent selection methods within a household. This method asks the informant to identify the household member that most recently had his or her birthday.

There are also two main theories that exist about the occurrence of overcoverage. The first theory is again based on the informant's misunderstanding of the definition of or rules about whom to count as a member of the sampling unit. An informant who does not understand the definition of who is a member of the sampling unit may also list people that do not belong to this sampling unit, and therefore someone might be listed twice in the sampling frame if the other unit to which the person belongs also is sampled. An example of this occurrence is a child that goes to college and does not live at home anymore but still is counted by his or her parents as being a member of the household. The second explanation for overcoverage problems is the difficulty in establishing the sampling frame for a fixed point in time to define the residency of all members of the sampling units selected in the survey. Generally, the problem of undercoverage bias seems to be more serious than the problem of overcoverage bias.

Sonja Ziniel

See also Informant; Last-Birthday Selection; Overcoverage; Residence Rules; Sampling Frame; Target Population; Undercoverage; Within-Unit Selection

Further Readings

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Groves, R.M. (1989). *Survey errors and survey costs.* New York: Wiley.
Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, *63*, 220–236.
Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? Survey undercoverage and household roster questions. *Journal of Official Statistics*, *13*, 1–18.

Within-Unit Selection

In survey research, a sample of households usually must be converted into a sample of individuals. This often is accomplished by choosing respondents within households using methods intended to yield a sample similar to the population of interest. Ideally, this is done with a probability technique because all unit members will have a known, nonzero chance of selection, thus allowing generalization to a population. Probability methods, however, tend to be time consuming and relatively intrusive because they ask about household composition, potentially alienating prospective respondents and therefore increasing nonresponse.

Most researchers have limited resources, so often they need quicker, easier, and less expensive quasi-probability or nonprobability methods that they believe will yield samples adequately resembling the population being studied. Although surveyors wish to minimize nonresponse and reduce noncoverage, they have to balance these choices to fit their goals and resources. They have to decide if the benefits of probability methods outweigh their possible contributions to total survey costs or if departures from probability selection will contribute too much to total survey error. This does not mean that they tolerate substandard practices but that they consider which trade-offs will be the most acceptable within their budget and time restrictions. One paramount goal is to gain the respondent's cooperation in as short a time as possible, and a second is to obtain a reasonable balance of demographic (usually age and gender) distributions. Most households are homogeneous in other demographic characteristics, such as education and race. Many well-respected polling and other survey research organizations have to make these kinds of choices.

Selection of respondents within households is particularly important in telephone surveys because most refusals occur at the inception of contact. Although it might seem advantageous to take the first person who answers the phone or to let interviewers choose respondents, those persons who are often the most available or most willing to be interviewed also tend to be disproportionately female or older, which may bias results. In addition, patterns of telephone answering are not random; they can vary by gender and by region. Researchers need to control respondent selection in systematic ways, therefore, even if methods are quasi-probability or nonprobability. The well-known Kish selection procedure is almost a pure probability technique. Interviewers list all adult males in the household and their relationships to others in order of decreasing age, then make a similar list of adult females. Interviewers then randomly select one person by consulting a set of tables. This technique is criticized as being time consuming for large households and potentially threatening to informants, especially women who may be concerned about their safety.

The "last-birthday," or "most recent birthday" method is a popular quasi-probability selection scheme, considered to be less intrusive and time consuming than the Kish method. Interviewers ask to speak to the adult in the household who had the last birthday. A variation is the less frequently used "next-birthday" method. In theory (but not necessarily in practice), the birthday methods are probability methods because they assume the first stage of a two-stage selection process is birth (expected to be a random event), and a second stage is selection into the sample.

To further streamline the selection process, V. Troldahl and R. E. Carter offered a nonprobability method requiring only two questions: the number of persons 18 years or older living in the household and the number of men. The number of tables for interviewers to consult shrank to four, calling for selection of the oldest man, the youngest man, the oldest woman, or the youngest woman. Barbara Bryant later proposed a modification that could better represent women. Others suggested asking for women rather than men, further altering what became known as the Troldahl-Carter-Bryant respondent selection method.

D. E. Hagan and C. M. Collier offered a further simplified plan with four forms that asked only for (a) the youngest man in the household, (b) the oldest man, (c) the youngest woman, and (d) the oldest woman. Forms A, B, and C were used two times out of seven, and form D was used one time in seven. If no such individual was in the household, the interviewer asked for the opposite sex of the same age group. This method was condensed even further into the youngest male/oldest female technique, in which the interviewer asks first for the youngest adult male. If there isn't one, the interviewer requests the oldest woman. Frequently, surveyors add "now at home" to improve response rates. Although these methods are intended to save time and to obtain age and sex distributions that approximate the general population, some researchers believe they distort distributions of gender within age or gender by household composition.

Louis Rizzo, J. Michael Brick, and Inho Park suggested a potentially shorter probability technique that is relatively nonintrusive and easy to implement in computer-assisted telephone interviewing random-digit dialing surveys. Interviewers need to know only the number of adults in one-adult or two-adult households, which make up about 85% of U.S. households. If the household is larger, the interviewer determines whether or not the informant is sampled. If not, another method, such as Kish or last birthday, is applied.

A number of studies have compared two or more different within-unit selection methods to aid researchers in decisions about procedures that will best fit their needs, although more research on these issues is desirable. All of these methods rely on the selection process to be done by an interviewer. Little research has been conducted on how to utilize a within-unit selection process when the survey is not administered by an interviewer.

Cecilie Gaziano

See also Computer-Assisted Telephone Interviewing (CATI); Hagan and Collier Selection Method; Kish Selection Method; Last-Birthday Selection; Noncoverage; Nonresponse; Troldahl-Carter-Bryant Respondent Selection Method; Within-Unit Coverage

Further Readings

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*, 124–157.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: Wiley.

Keeter, S., & Fisher, K. (1997–1998). Comparison of two respondent selection methods. *NNSP Newsletter*, *31*(Winter), 1–3.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Newbury Park, CA: Sage.

Rizzo, L., Brick, J. M., & Park, I. (2004). A minimally intrusive method for sampling persons in random digit dial surveys. *Public Opinion Quarterly*, *68*, 267–274.

Troldahl, V., & Carter, R. E., Jr. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, *1*, 71–76.

Voss, D. S., Gelman, A., & King, G. (1995). Preelection survey methodology: Details from eight polling organizations, 1988 and 1992. *Public Opinion Quarterly*, *59*, 98–132.

World Association for Public Opinion Research (WAPOR)

The World Association for Public Opinion Research (WAPOR) is an international organization of individual members who are engaged in public opinion research and the development of survey research methods. It was founded in 1947 to promote and improve the quality of these kinds of activities around the world. WAPOR collaborates with both the American Association for Public Opinion Research (AAPOR) and the European Society for Opinion and Marketing Research (ESOMAR), with which it respectively shares its annual conferences every other year. This typically means that in alternating years, the WAPOR conference is held in North America and then in Europe. WAPOR is also affiliated with the United Nations' Educational, Scientific and Cultural Organization (UNESCO) through which it promotes a variety of social science efforts and projects. Members currently come from almost 60 different countries around the world.

WAPOR organizes a variety of activities to promote quality research around the world. It sponsors two or three regional seminars at different venues each year that focus on research methods, specific uses of public opinion research related to elections and democracy, and survey data quality. Recent regional conferences have been held in Latin America (Uruguay), the Middle East (Israel), Asia (India), and Europe (Italy). WAPOR is especially interested in promoting the training of journalists and others so that they can better report on public opinion. The organization has a long-standing interest in maintaining freedom to publish the results of public opinion polls, and it actively responds to legislation or attempts to legislate restrictions on the dissemination of such information. It disseminates a report on this topic, titled *Freedom to Publish Opinion Polls*, which is updated periodically to indicate where such restrictions are, and are not, in place.

In conjunction with ESOMAR, WAPOR promotes the *Guide to Opinion Polls*, which includes an international code of practice for publication of public opinion poll results. In recent years, the organization has turned its attention to setting broad standards for widely used or potentially adoptable public opinion methodologies. This began with the development of WAPOR Guidelines for Exit Polls and Election Forecasts, a set of desirable procedures for the conduct and dissemination of information about election results involving interviews with voters who have just cast ballots. It is also at work to develop a similar set of guidelines and standards for the conduct of "peace polls," studies of opinions about the sources and bases of conflict in locations around the world, including the need for wide dissemination of such information, through the media, to all affected communities.

The *International Journal of Public Opinion Research*, WAPOR's quarterly journal, publishes timely research on public opinion, especially in comparative perspective, and on research methodology. The content also includes summaries of recent public opinion findings and results published in books and other venues, as well as book reviews. WAPOR also produces a quarterly newsletter.

WAPOR is governed by an 11-person executive council, 5 of whom are elected by the membership at large; the remaining 6 council members are appointed. The executive council meets between annual conferences as necessary to conduct its business. The dues structure in WAPOR employs three tiers to take into account different national economic conditions around the world, and there is a special reduced fee for students.

Michael W. Traugott

See also American Association for Public Opinion Research (AAPOR); Election Night Projections; Exit Polls; *International Journal for Public Opinion Research* (IJPOR); Prior Restraint

Further Readings

European Society for Opinion and Marketing Research (ESOMAR). (n.d.). *ESOMAR/WAPOR Guide to Opinion Polls including the ESOMAR International Code of Practice for the Publication of Opinion Polls*. Retrieved April 7, 2008, from http://www.esomar.org/uploads/pdf/ESOMAR_Codes&Guidelines_OpinionPolling_v5.pdf

World Association for Public Opinion Research: http://www.unl.edu/WAPOR/index.htm

Z

ZERO-NUMBER BANKS

Telephone numbers in the United States consist of 10 digits: The first three are the area code; the next three are the prefix or exchange; the final four digits are the suffix or local number. The 10,000 possible numbers for a suffix can be subdivided into banks of consecutive numbers: 1,000-banks (Nnnn); 100-banks (NNnn); or 10-banks (NNNn). Zero-number banks, or zero-listed banks, are banks that do not contain directory-listed residential numbers. Although including zero-number banks in a random-digit dialing frame allows for 100% coverage of landline residential telephone numbers, their inclusion can substantially reduce sample efficiency.

Based on regular analyses conducted by Survey Sampling International, only 29% of 100-banks in POTS (Plain Old Telephone Service) prefixes, and other prefixes shared by different types of service, have at least one directory-listed number. In prefixes that have at least one directory-listed telephone number, only 47% of the possible 100-banks contain at least one directory-listed number. Because of these inefficiencies, most random-digit dialing surveys today use list-assisted, truncated frames—that is, frames truncated to include only listed-banks or those 100-banks that contain at least one directory-listed residential telephone number.

This truncation introduces some coverage error by excluding 100-banks that contain residential numbers but are missing from the truncated frame. The most common reason for this omission is that newly opened banks have not yet been published in a telephone directory. Because directories are published only once a year, the time lag between number assignment and directory publication can result in new blocks not being represented. Another source of error is common in rural areas, particularly those serviced by small, local telephone companies. A formal directory may not be readily available to compilers, but numbers are listed in a paperback book that looks like the local real estate listings available at the supermarket.

Alternative sample designs are available for researchers that opt to include zero-listed banks. One approach, the Mitofsky-Waksberg method, takes advantage of the tendency of telephone numbers to cluster in 100-banks. It starts with a sample of primary numbers in prefixes available for landline residential use. If a primary number is determined to be a working residential number, a cluster of additional numbers in generated in the same 100-bank. Another approach is to use disproportionate stratified samples of both zero-listed banks and listed banks. For many years, this design was the sampling protocol for all surveys conducted as part of the U.S. Department of Health Behavioral Risk Factor Surveillance System.

Research by Michael Brick and others suggests that this coverage error is 3–4% of telephone households. However, work by Brick and by Clyde Tucker and Jim Lepkowski confirms that the efficiency gains of list-assisted designs make them preferable in most cases. In fact, in 2003 the Behavioral Risk Factor Surveillance System protocol was revised to include only the listed-bank stratum.

Linda Piekarski

See also Behavioral Risk Factor Surveillance System (BRFSS); Coverage; Coverage Error; Mitofsky-Waksberg Sampling; Random-Digit Dialing (RDD); Suffix Banks

Further Readings

Brick, J. M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, *59*(2), 218–235.

District of Columbia Department of Health. (n.d.). *BRFSS survey protocol.* Retrieved December 2, 2006, from http://www.dchealth.dc.gov/doh/cwp/view,a,1374,q,601933.asp

Lepkowski, J. M. (1988). Telephone sampling methods in the United States. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 73–98). New York: Wiley.

National Center for Chronic Disease Prevention and Health Promotion. (2003). *BRFSS policy memo 2003.1: Sampling and sampling information protocols.* Retrieved December 2, 2006, from http://www.cdc.gov/brfss/technical_infodata/memos/20031.htm

Tucker, C., Lepkowski, J. M., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly*, *66*, 321–338.

z-Score

The *z*-score is a statistical transformation that specifies how far a particular value lies from the mean of a normal distribution in terms of standard deviations. *z*-scores are particularly helpful in comparing observations that come from different populations and from distributions with different means, standard deviations, or both. A *z*-score has meaning only if it is calculated for observations that are part of a normal distribution.

z-scores are sometimes referred to as *standard scores.* When the values of a normal distribution are transformed into *z*-scores, the transformed distribution is said to be "standardized" such that the new distribution has a mean equal to 0 and a standard deviation equal to 1.

The *z*-score for any observation is calculated by subtracting the population mean from the value of the observation and dividing the difference by the population standard deviation, or $z = (x - \mu)/\sigma$. Positive *z*-scores mean that the observation in question is greater than the mean; negative *z*-scores mean that it is less than the mean. For instance, an observation with a *z*-score of 1.0 would mean that the observation is exactly one standard deviation above the mean of the distribution. An observation with a *z*-score equal to –0.5 would fall one-half of one standard deviation below the distribution's mean. An observation with a *z*-score equal to 0 would be equal to the mean of the distribution.

As an example, a researcher looking at middle school students' test scores might benefit from using *z*-scores as a way of comparing the relative performance of a seventh-grade student on a seventh-grade test to an eighth-grade student on an eighth-grade test. In this example, the researcher knows that the scores for the entire population of seventh graders and for the entire population of eighth graders are normally distributed. The average number of correct answers (out of 100 multiple choice questions) for the population of seventh graders on the seventh-grade test is 65 with a standard deviation of 10. The average score (out of 100 multiple choice questions) for the population of eighth graders on the eighth-grade test is 72 with a standard deviation of 12.

The seventh- and eighth-grade students of interest to this researcher scored 70 correct and 75 correct, respectively. Transforming each raw score into a *z*-score would be an appropriate way to determine which student scored better relative to his or her own population (cohort). The *z*-score for the seventh-grade student would be (70 – 65)/10, or 0.5, meaning that he or she scored 0.5 standard deviation above the average for seventh-grade students. The *z*-score for the eighth-grade student would be (75 – 72)/12, or 0.25, meaning that he or she scored 0.25 standard deviation above the average for eighth-grade students. Relative to his or her peers, the seventh-grade student performed better than the eighth-grade student, despite the eighth grader's higher raw score total.

Joel K. Shapiro

See also Percentile; Population of Interest; Population Parameter; Raw Data

Further Readings

Kvanli, A. H., Guynes, C. S., & Pavur, R. J. (1986). *Introduction to business statistics.* St. Paul, MN: West Publishing.

Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics* (5th ed.). New York: Wiley.

Index

Entry titles and their page numbers are in bold; illustrative material is identified by (figure) and (table).