STATISTICAL POLICY

WORKING PAPER 31

Measuring and Reporting Sources of Error in Surveys

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

July 2001

The Federal Committee on Statistical Methodology
April 2001

MEMBERS

Virginia de Wolf, Acting Chair
Committee on National Statistics

Susan W. Ahmed
Consumer Product Safety Commission

Wendy L. Alvey, Secretary
U.S. Bureau of the Census

Lynda Carlson
National Science Foundation

Cynthia Z.F. Clark
U.S. Bureau of the Census

Steven B. Cohen
Agency for Healthcare Research and Quality

Lawrence H. Cox
National Center for Health Statistics

Cathryn Dippo
U.S. Bureau of Labor Statistics

Zahava D. Doering
Smithsonian Institution

Robert E. Fay
U.S. Bureau of the Census

Ronald Fecso
National Science Foundation

Gerald Gates
U.S. Bureau of the Census

Barry Graubard
National Cancer Institute

William Iwig
National Agricultural Statistics Service

Daniel Kasprzyk
National Center for Education Statistics

Nancy J. Kirkendall
Energy Information Administration

Charles P. Pautler, Jr.
Internal Revenue Service

Susan Schechter
U.S. Office of Management and Budget

Rolf R. Schmitt
Federal Highway Administration

Monroe G. Sirken
National Center for Health Statistics

Nancy L. Spruill
U.S. Department of Defense

Clyde Tucker
U.S. Bureau of Labor Statistics

Alan R. Tupek
U.S. Bureau of the Census

Denton R. Vaughan
U.S. Bureau of the Census

G. David Williamson
Centers for Disease Control and Prevention

Alvan O. Zarate
National Center for Health Statistics

CONSULTANT

Robert Groves
Joint Program in Survey Methodology

STATISTICAL POLICY
WORKING PAPER 31

Measuring and Reporting Sources of Error in Surveys

Prepared by the
Subcommittee on Measuring and Reporting the Quality of Survey Data

Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

June 2001

Members of the Subcommittee on Measuring and Reporting the Quality of Survey Data

Daniel Kasprzyk, Chair
National Center for Education Statistics
(Education)

Dale Atkinson
National Agricultural Statistics Services
(Agriculture)

Judith Conn
Centers for Disease Control and Prevention
(Health and Human Services)

Ram Chakrabarty[1]
Bureau of the Census
(Commerce)

Charles Darby
Agency for Healthcare Research and Quality

Lee Giesbrecht[2]
Bureau of Transportation Statistics
(Transportation)

Brian Harris-Kojetin[3]
Bureau of Labor Statistics
(Labor)

Howard Hogan
Bureau of the Census
(Commerce)

Nancy Kirkendall
Energy Information Administration
(Energy)

Marilyn McMillen
National Center for Education Statistics
(Education)

Renee Miller
Energy Information Administration
(Energy)

Chris Moriarity
National Center for Health Statistics
(Health and Human Services)

Dennis Schwanz
Bureau of the Census
(Commerce)

Carolyn Shettle[4]
National Science Foundation

W. Karl Sieber
National Institute for Occupational Safety
and Health
(Health and Human Services)

Antoinette Ware-Martin
Energy Information Administration
(Energy)

John Wolken
Federal Reserve Board

Graham Kalton, Senior Advisor
Joint Program in Survey Methodology and
WESTAT

[1] Deceased

[2] Now with Abt Associates, Inc.

[3] Now with Arbitron, Inc.

[4] Now with Institute for Survey Research, Temple University

Acknowledgments

In 1996, Maria Gonzales, Chair of the Federal Committee on Statistical Methodology (FCSM), formed the Subcommittee on Measuring and Reporting the Quality of Survey Data. Members of the subcommittee were drawn from 12 federal agencies whose missions include the collection, production, dissemination, and reporting of statistical information.

Maria's energy and enthusiasm provided the early stimuli for the subcommittee, and her untimely death was a loss to the subcommittee. A little over one year after Maria's death, the subcommittee renewed its commitment to the project. Nancy Kirkendall, then Chair of the FCSM, and Dan Kasprzyk, Chair of the FCSM Subcommittee on Measuring and Reporting the Quality of Survey Data, in consultation with Graham Kalton identified broad areas for the subcommittee to consider. The members of the subcommittee, drawing on their own experience as well as the culture and norms of their respective agencies, took those broad ideas and developed the approach found in this working paper.

The working paper is the result of several activities. First, many discussions and meetings of the subcommittee occurred over the course of several years. These meetings provided the opportunity to establish common ground and shared interests among the members of the subcommittee. Second, members of the subcommittee organized sessions related to the topic of "measuring and reporting the quality of survey data" at the 1996 and 1998 Federal Committee on Statistical Methodology (FCSM)/Council on Professional Associations for Federal Statistics (COPAFS) conferences. The papers presented at these conferences are available in the conference proceedings (OMB Statistical Policy Working Paper 26: *Seminar on Statistical Methodology in the Public Service* and OMB Statistical Policy Working Paper 28: *Seminar on Interagency Coordination and Cooperation*). Third, the 1997 Statistics Canada Symposium, "New Directions in Surveys and Censuses," provided an opportunity to present a paper on the general approach taken by the subcommittee and obtain comments from the statistical community. Fourth, the subcommittee conducted three studies to develop an understanding and appreciation of the reporting practices of Federal statistical agencies. Finally, the subcommittee organized an invited session for the 1999 International Statistical Institute meetings in Helsinki, Finland, at which the international statistics community commented on the results of the three studies.

These activities helped guide the development and are the basis of this working paper. All subcommittee members participated in spirited, informative, and productive discussions over the course of several years. Video conferencing through the auspices of the National Center for Health Statistics allowed full participation of subcommittee members not located in the Washington, DC metropolitan area.

Each chapter in this report was drafted by one or more members of the subcommittee as follows:

Chapter	Author
1	Daniel Kasprzyk
2	Daniel Kasprzyk
3	Chris Moriarity
4	Marilyn McMillen, Brian Harris-Kojetin, Renee Miller, and Antoinette Ware-Martin
5	Howard Hogan and Karl Seiber
6	Dennis Schwanz, Charles Darby, and Judith Conn
7	Dale Atkinson and John Wolken
8	Pat Dean Brick, Lee Giesbrecht, and Carolyn Shettle

Jim Knaub (Energy Information Administration) and the late Ram Chakrabarty (U.S. Bureau of the Census) contributed to the early discussions of the subcommittee. Lee Giesbrecht, Carolyn Shettle, and Brian Harris-Kojetin were active subcommittee members, drafting major parts of individual chapters, until each left government service for the private sector in 2000, 1999, and 1998, respectively. Pat Dean Brick, a consultant now with Research Triangle Institute, provided substantial editorial assistance on the entire manuscript. Daniel Kasprzyk was responsible for the final edit and review of the manuscript.

The subcommittee appreciates the thoughtful and helpful comments of the reviewers of the working paper: Cynthia Clark, Suzanne Dorinski, Carma Hogue, Susan Schecter, Clyde Tucker, and Al Tupek. Throughout the subcommittee's existence, Graham Kalton provided timely and practical advice. His experience, knowledge and good judgment helped clarify issues for the subcommittee; his commentary on interim products and his dedication to producing outstanding work is appreciated. Nancy Kirkendall, as chair of the FCSM, encouraged and cajoled the subcommittee in its efforts. She was an active subcommittee member and made substantial content and editorial contributions to all chapters. Finally, the subcommittee thanks Carol Rohr and Susan Baldridge of Pinkerton Computer Consultants, Inc. for their excellent effort in editing and desktop publishing the manuscript.

Summary

In 1996, the FCSM established a subcommittee to review the measurement and reporting of data quality in federal data collection programs. Many issues revolve around these two broad topics, not the least of which is what is meant by "quality." Different data users have different goals and, consequently, different ideas of what constitutes "quality." If defining quality is difficult, then the reporting of quality is also. Reporting "quality" is dependent on the needs of data users and the kind of product—analytic report, technical report or data set, for example, made available to the user. The FCSM subcommittee, whose membership represents the experiences of 12 statistical agencies, took the approach of studying "data quality" in terms of the measurement and reporting of various error sources that affect data quality: sampling error, nonresponse error, coverage error, measurement error, and processing error.

The subcommittee developed an approach to studying these error sources by trying to answer four questions:

- What measurement methods are used by federal data collection programs to assess sources of error in data collection programs?
- To what extent do federal data collection programs report information on sources of error to the user community?
- How does reporting about error sources vary across different types of publications and dissemination media?
- What information on sources of error should federal data collection programs provide and how should they provide it?

This report represents the subcommittee's efforts at addressing these questions. To understand current reporting practices, the subcommittee conducted three studies. Two studies focussed on specific kinds of reports—the short-format report and the analytic report. The third study reviewed the extent to which information about error sources was available on the Internet. The studies' results were surprising because information about sources of survey error were not as well-reported as the subcommittee had expected. Chapter 1 discusses data quality and policies and guidelines with respect to reporting on the quality of data. Chapter 2 describes the studies, the studies' results and provides recommendations on the kind of information that ought to be reported in short-format reports, analytic reports, and technical reports. Chapters 3–7 describe the measurement of sources of error in surveys: sampling error, nonresponse error, coverage error, measurement error, and processing error. Each chapter discusses the results of the subcommittee's studies that relate to the particular source of error, provides specific recommendations for reporting error sources in an analytic report, and identifies additional topics to report in the technical report format. Chapter 8 discusses the measurement and reporting of total survey error.

Contents

List of Tables

Chapter 1

Measuring and Reporting Data Quality in Federal Data Collection Programs

1.1 Introduction

The United States statistical system includes over 70 statistical agencies spread among 10 separate departments and 8 independent agencies. Within this decentralized statistical system the Office of Management and Budget (OMB) plays an important leadership role in the coordination of statistical work across the federal government. For example, the Federal Committee on Statistical Methodology (FCSM), a committee of the OMB, has played a leadership role in discussions of the methodology of federal surveys (Gonzalez 1995; Bailar 1997) for almost 25 years.

In 1996, the FCSM established a subcommittee to review the measurement and reporting of data quality in federal data collection programs. The issues contained within this broad mandate are complex. Measuring the quality of survey data takes on different meanings depending on the constituency. Different data users have different goals and, consequently, different ideas of what constitutes quality. Similarly, the reporting of "quality" can be implemented quite differently depending on the type of data product produced. The FCSM subcommittee, whose membership represents the experiences of twelve statistical agencies, developed an approach to examining this topic by trying to provide answers to four questions:

- What measurement methods are used by federal agencies to assess sources of error in data collection programs?

- To what extent do federal data collection programs report information on sources of error to the user community?

- How does reporting about error sources vary across different types of publications and dissemination media?

- What information on sources of error should federal data collection programs provide and how should they provide it?

This report represents the subcommittee's efforts at addressing these questions. In general, the subcommittee took the approach of studying data quality in terms of the measurement and reporting of various error sources that affect data quality: sampling error, nonresponse error, coverage error, measurement error, and processing error (see, for example Kish 1965). The result of this analysis forms the core of this report.

1.2 Overview of the Report

This report provides a general discussion of sources of error in data collection programs. Chapter 2 describes studies undertaken by the subcommittee to understand current statistical agency practices on the reporting of those sources of error, and it gives general recommendations on the types of information about those error sources that ought to be available in short-format reports, analytic reports, or on the Internet. Each of chapters 3–7 addresses a specific source of error from two directions: first, the measurement of the source of error. In this case, a brief discussion of the methods used to measure the error source is given. The second direction is the nature and extent of reporting sources of error in analytic applications and more comprehensive survey design and methodological reports. Examples of practice, both in measuring the error source as well as how the error source is reported, are given when it is appropriate. The final chapter, chapter 8, discusses total survey error, the ways in which it is measured and reported by federal statistical agencies, and several recommendations concerning the reporting of total survey error.

1.3 Data Quality

A rich literature exists and continues to grow on the topic of survey data quality (Lyberg et al. 1997; Collins and Sykes 1999) and its management in national statistical agencies (Brackstone 1999). Definitions of the concept proliferate, but cluster around the idea that the characteristics of the product under development meet or exceed the stated or implied needs of the user. Arondel and Depoutot (1998) suggest in their review that statistical organizations should break down quality into components or characteristics that focus around several key concepts: accuracy, relevance, timeliness, and accessibility. See also, Statistics Canada (1992) and Statistics Sweden (Andersson, Lindstrom, and Lyberg 1997).

Accuracy is an important and visible aspect of quality that has been of concern to statisticians and survey methodologists for many years. It relates to the closeness between estimated and true (unknown) values. For many, accuracy means the measurement and reporting of estimates of sampling error for sample survey programs, but, in fact, the concept is much broader, taking in nonsampling error as well. Nonsampling error includes coverage error, measurement error, nonresponse error, and processing error. These sources of error will be discussed below; however, it is important to recognize that the accuracy of any estimate is affected by both sampling and nonsampling error.

Relevance refers to the idea that the data collection program measures concepts that are meaningful and useful to data users. Does the concept implemented in the data collection program fit the intended use? For example, concepts first measured in a continuous sample survey program 20 years ago may be inapplicable in current society; that is, it may no longer be relevant to data users. Determining the relevance of concepts and definitions is a difficult and time-consuming process requiring the expertise of data collectors, data providers, data users, agency researchers, and expert panels.

Timeliness can refer to several concepts. First, it refers to the length of the data collection's production time—the time from data collection until the first availability of a product. Fast release times are without exception looked upon favorably by end users. Second, timeliness can also refer to the frequency of the data collection. Timely data are current data. Timeliness can be difficult to characterize since the characteristics of the data collection can affect the availability of data. For example, a new sample survey may require more time prior to implementation than the revision of an existing survey. Data from continuous recurring surveys should be available sooner than periodic or one-time surveys, but ultimately timeliness is assessed by user needs and expectations.

Accessibility, as a characteristic of data quality, refers to the ability of data users to obtain the products of the data collection program. Data products have their most value—are most accessible—when they are easily available to end-users and in the forms and formats desired. Data products are of several types—individual microdata in user-friendly formats on different media, statistical tabulations on key survey variables, and analytic and descriptive analysis reports. Accessibility also implies the data products include adequate documentation and discussion to allow proper interpretation of the survey results. Accessibility can also be described in terms of the efforts data producers make to provide "hands-on" technical assistance in using and interpreting the data products through consultation, training classes, etc.

Arondel and Depoutot (1998) suggest three other characteristics of data quality: comparability of statistics, coherence, and completeness. Comparability of statistics refers to the ability to make reliable comparisons over time; coherence refers to the ability of the statistical data program to maintain common definitions, classifications, and methodological standards when data originate from several sources; and completeness is the ability of the statistical data collection to provide statistics for all domains identified by the user community.

Survey data quality is a concept with many dimensions and with each dimension linked with others. In the abstract, all dimensions of data quality are very important, but in practice, it is usually not possible to place high importance on all dimensions. Thus, with fixed financial resources, an emphasis on one dimension will result in a decrease in emphasis in another. More emphasis on accuracy can lead to less emphasis on timeliness and accessibility; or an emphasis on timeliness may result in early/preliminary release data of significantly lower accuracy. Each dimension is important to an end user, but each user may differ in identifying the most important priorities for a data collection program.

The subcommittee chose to limit its coverage to the accuracy dimension—a dimension that has a history of measurement and reporting. The subcommittee focused on reviewing statistical indicators used to describe different aspects of survey accuracy in relation to various error sources, how indicators may be measured, and whether and how they are presented to data users.

1.4 Data Quality Policies and Guidelines

1.4.1 The Principle of Openness

The subcommittee's focus on accuracy is rooted in two long-standing important general principles/features of government statistical systems. The first and critical feature of federal statistical agencies is that there is a stated policy of "openness" concerning the reporting of data quality to users. The U.S. Office of Management and Budget (1978) states that:

> "To help guard against misunderstanding and misuse of data, full information should be available to users about sources, definitions, and methods used in collecting and compiling statistics, and their limitations."

Indeed, the principle is characteristic of national statistical systems; for example, Statistics Canada (1992) has articulated a policy on informing users about data quality and methodology:

> “Statistics Canada, as a professional agency in charge of producing official statistics, has the responsibility to inform users of concepts and methodology used in collecting and processing its data, the quality of the data it produces, and other features of the data that may affect their use and interpretation.”

New Zealand (Statistics New Zealand 1998) has developed protocols to guide the production and release of official statistics. The protocols are based on ten principles—one of which is to

> “Be open about methods used and documentation of methods and quality measures should be easily available to users to allow them to determine fit for use.”

The policy of openness in providing full descriptions of data, methods, assumptions, and sources of error is one that is universally accepted by United States statistical agencies as well as the statistical agencies of other countries. As a policy and as a characteristic of a government statistical system, it is noncontroversial. There is general agreement that data users need information on the sources and methods used to compile the data. They also need to understand the nature and sources of error in sample surveys. Correct interpretation or re-analysis of data relies on the availability of such information. As Citro (1997) points out, data users must have the opportunity to review data methods and data limitations; they cannot heed these limitations if they do not have the opportunity to study and review them. However, as will be discussed below, implementation of the policy can vary in many ways.

1.4.2 Statistical Standards and Guidelines

The second feature of the United States statistical system that is important is the availability of standards or guidelines for survey processes. The U.S. Office of Management and Budget (1978) provides general guidelines for reporting survey information. Individual agencies have the flexibility to develop the general guidelines into more specific guidelines that help the agency codify its own professional standards. The specific guidelines help to promote consistency among studies, and promote the documentation of methods and principles used in collection, analysis, and dissemination. These standards and guidelines generally include prescriptions for the dissemination of information about data quality to users. Such guidelines are not uniformly available across all agencies, but a few good examples exist, such as the standards developed by the U.S. Department of Energy (1992) and the U.S. Department of Education (1992). The standards and guidelines have the effect of helping staff improve the quality and uniformity of data collection, analysis, and dissemination activities.

Other agencies take a slightly different approach to the development of standards and guidelines, focussing more on the establishment of policies and procedures for reviewing reports, determining sampling errors, and testing hypotheses (Sirken et al. 1974) or presenting information concerning sampling and nonsampling error (Gonzalez et al. 1975; updated by Bailar 1987).

The interest in establishing standards and guidelines is also found in the statistical systems of other countries. For example, the United Kingdom Government Statistical Service (1997) has developed guidelines that focus specifically on the reporting of data quality. These guidelines are in the form of a checklist of questions related to individual areas of the survey process. Statistics Canada (1998) has developed a report that provides "good practices" in the form of principles and guidelines for the individual steps of a survey. Documenting data quality in relation to various error sources is not a new concern. Thirty-seven years ago, the United Nations (1964) presented recommendations on the topics to be documented when preparing sample survey reports, including information on many sources of error. The United Nations recommendations include the provision of detailed information about how the survey was conducted.

Information about survey procedures provides users with valuable insights into data quality. As an example, in a face-to-face household survey, information about interviewer training and recruitment, the extent of checking on the conduct of interviews, the number of visits required, and the refusal conversion procedures is useful in assessing the quality of the conduct of the survey. Therefore, it is important for data producers to report information about the background and history of the data collection program, its scope and objectives, sample design and sample size, data collection procedures, time period of data collection, the response mode, the designated respondents, as well as processing and estimation procedures. For repeated surveys, it is important for users to be aware of changes in design, content, and implementation procedures since they may affect comparisons over time.

1.4.3 Discussion

"Openness" and "measurement and reporting guidelines"—two principles discussed above—provide the impetus for the work of the subcommittee. The principle of "openness" is important for the United States statistical system because it helps others to reproduce results and question findings. The principle allows the system and the agencies in it to be held accountable and to maintain impartiality for the presentation and reporting of official statistics. The second principle provides the policies and guidelines to implement the policy of openness.

The measurement and reporting of error sources is important for everyone who uses statistical data. For the analyst, this information helps data analyses through an awareness of the limitations of the data. It helps the methodologist understand current data collection procedures, methods, and data limitations, and it motivates the development of better methods for future data collections. For the statistical agency, the implementation of effective measurement and reporting is an integral part of the good practices expected of a statistical agency.

1.5 Measuring Sources of Error

The subcommittee organized its work in terms of the error sources that affect accuracy. It reviewed methods for measuring error sources and reviewed indicators for describing information on data quality. The subcommittee identified five sources of error: sampling error, nonresponse error, coverage error, measurement error, and processing error.

Sampling error is probably the best-known source of survey error and refers to the variability that occurs by chance because a sample rather than an entire population was surveyed. The reporting of sampling error for survey estimates should be important to all statistical agencies. For any survey based on a probability sample, data from the survey can be used to estimate the standard errors of survey estimates. Nowadays, the standard errors for most estimates can be readily computed using software that takes into account the survey's complex sample design. The challenge that occurs with the computation of standard errors is a result of the multi-purpose nature of many federal surveys. Surveys produce many complex statistics and the task of computing and reporting standard errors for all the survey estimates and for differences between estimates is an extremely large one.

Nonresponse error is a highly visible and well-known source of nonsampling error. It is an error of nonobservation reflecting an unsuccessful attempt to obtain the desired information from an eligible unit. Nonresponse reduces sample size, results in increased variance, and introduces a potential for bias in the survey estimates. Nonresponse rates are frequently reported and are often viewed as a proxy for the quality of a survey. Nonresponse rates may be calculated differently for different purposes (Lessler and Kalsbeek 1992; Gonzalez, Kasprzyk, and Scheuren 1994; Council of American Survey Research Organizations 1982; American Association for Public Opinion Research 2000) and they are often miscalculated. The complexities of the survey design often make calculation and communication of response rates confusing and potentially problematic. While reporting nonresponse rates is important, nonresponse rates alone provide no indication of nonresponse bias. Special studies are necessary.

Coverage error is the error associated with the failure to include some population units in the frame used for sample selection (undercoverage) and the error associated with the failure to identify units represented on the frame more than once (overcoverage). The source of coverage error is the sampling frame itself. It is important, therefore, that information about the quality of the sampling frame and its completeness for the target population is known. Measurement methods for coverage error rely on methods external to the survey operations; for example, comparing survey estimates to independent sources or by implementing a case-by-case matching of two lists.

Measurement error is characterized as the difference between the observed value of a variable and the true, but unobserved, value of that variable. Measurement error comes from four primary sources in survey data collection: the questionnaire, as the official presentation or request for information; the data collection method, as the way in which the request for information is made; the interviewer, as the deliverer of the questions; and the respondent, as the recipient of the request for information. These sources comprise the entirety of data collection, and each source can introduce error into the measurement process. For example, measurement error may occur in respondents' answers to survey questions, including misunderstanding the meaning of the question, failing to recall the information accurately, and failing to construct the response correctly (e.g., by summing the components of an amount incorrectly). Measurement errors are difficult to quantify, usually requiring special, expensive studies. Reinterview programs, record check studies, behavior coding, cognitive testing, and randomized experiments are a few of the approaches used to quantify measurement error.

Processing error occurs after the survey data are collected, during the processes that convert reported data to published estimates and consistent machine-readable information. Each processing step, from data collection to the publication of the final survey results, can generate errors in the data or in the published statistics. These errors range from a simple recording error, that is a transcribing or transmission error, to more complex errors arising from a poorly specified edit or imputation model. They tend not to be well-reported or well-documented, and are seldom treated in the survey research literature. Processing errors include data entry, coding, and editing and imputation errors. Imputation errors are included under processing error because many agencies treat failed edits as missing and impute values for them. Error rates are determined through quality control samples; however, in recent years authors have advocated continuous quality management practices (Morganstein and Marker 1997; Linacre and Trewin 1989).

The classification of error sources in surveys described above provides a framework for users of statistical data to develop an understanding of the nature of the data they analyze. An understanding of the limitations of data can assist an analyst in developing methods to compensate for the known shortcomings of their data. Of course, the errors from various sources are not of the same size or of the same importance. Later chapters will describe measurement techniques for determining the magnitude of the sources of error.

References

American Association for Public Opinion Research. 2000. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* Ann Arbor, MI: AAPOR.

Andersson, C., Lindstrom, H., and Lyberg, L. 1997. "Quality Declaration at Statistics Sweden." *Seminar on Statistical Methodology in the Public Service.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 26). 131–144.

Arondel, P. and Depoutot, R. May 1998. "Overview of Quality Issues when Dealing with Soci-economic Products in an International Environment." Paper prepared for presentation at the XXXth ASU Meeting.

Bailar, B. 1997. "The Federal Committee on Statistical Methodology." *Proceedings of the Section on Government Statistics and Section on Social Statistics.* Alexandria, VA: American Statistical Association. 137–140.

Bailar, B. June 2, 1987. "Policy on Standards and Review of Census Bureau Publications." U.S. Bureau of the Census memorandum.

Brackstone, G. 1999. "Managing Data Quality in a Statistical Agency." *Survey Methodology.* 25(2): 139–149.

Citro, C. 1997. "Discussion." *Seminar on Statistical Methodology in the Public Service.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 26). 43–51.

Collins, M. and Sykes, W. 1999. "Extending the Definition of Survey Quality." *Journal of Official Statistics.* 15(1): 57–66.

Council of American Survey Research Organizations. 1982. *On the Definitions of Response Rates.* Port Jefferson, NY.

Gonzalez, M.E. 1995. "Committee Origins and Functions: How and Why the Federal Committee on Statistical Methodology Began and What it Does." *Proceedings of the Section on Government Statistics.* Alexandria, VA: American Statistical Association. 262–267.

Gonzalez, M.E., Kasprzyk, D., and Scheuren, F. 1994. "Nonresponse in Federal Surveys: An Exploratory Study." *Amstat News* 208. Alexandria, VA: American Statistical Association.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. 1975. "Standards for Discussion and Presentation of Errors in Survey and Census Data." *Journal of the American Statistical Association.* 70(351)(II): 5–23.

Kish, L. 1965. *Survey Sampling.* New York: John Wiley & Sons.

Lessler, J.T. and Kalsbeek, W.D. 1992. *Nonsampling Error in Surveys.* New York: John Wiley & Sons.

Linacre, S. and Trewin, D. 1989. “Evaluation of Errors and Appropriate Resource Allocation in Economic Collections.” *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census. 197–209.

Lyberg, L., Biemer, P., Collins, M., deLeeuw E., Dippo, C., Schwarz, N. and Trewin, D. (eds.) 1997. *Survey Measurement and Process Quality.* New York: John Wiley & Sons.

Morganstein, D. and Marker, D. 1997. “Continuous Quality Improvement in Statistical Agencies.” In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality.* New York: John Wiley & Sons. 475–500.

Sirken, M. G., Shimizu, B.I., French, D.K., and Brock, D.B. 1974. *Manual on Standards and Procedures for Reviewing Statistical Reports.* Washington, DC: National Center for Health Statistics.

Statistics Canada. 1998. *Statistics Canada Quality Guidelines.* Ottawa, Canada.

Statistics Canada. 1992. *Policy on Informing Users of Data Quality and Methodology.* Ottawa, Canada.

Statistics New Zealand. August 1998. *Protocols for Official Statistics.* Statistics New Zealand.

United Kingdom Government Statistical Service. 1997. *Statistical Quality Checklist.* London: U.K. Office for National Statistics.

United Nations. 1964. *Recommendations for the Preparation of Sample Survey Reports (Provisional Issue).* Statistical Papers, Series C, ST/STAT/SER.C/1/Rev.2. New York: United Nations.

U.S. Department of Education. 1992. *NCES Statistical Standards.* Washington, DC: National Center for Education Statistics (NCES 92–021r).

U.S. Department of Energy. 1992. *The Energy Information Administration Standards Manual.* Washington DC: Energy Information Administration.

U.S. Office of Management and Budget. 1978. *Statistical Policy Handbook.* Washington, DC.

Chapter 2

Reporting Sources of Error: Studies and Recommendations

2.1 Reporting Formats and Reporting Sources of Error

Reporting formats for presenting information to users vary considerably not only across statistical agencies but also within a single statistical agency. Individual data programs have a variety of constituencies and user groups, each with diverse representation ranging from sophisticated data analysts with graduate degrees to reporters and the general public. These user groups are served by different types of data products. Individual-level data sets provide survey responses in a format that can be accessed by data analysts. Data sets are often packaged on CD-ROMs with software and instructions on how to create analytic subfiles in the formats used by statistical software. Other CD-ROM products may allow the tabulation of a large number (but not all) of survey variables. Lately, microdata are being made available on the Internet in the form of downloadable files and through online statistical tools. In contrast, diskettes and even 9-track tape files are still released to the general public.

Print products vary widely in their sophistication and complexity of analysis. Simple categorizations of print reports are difficult, but several broad types of reports can be distinguished. Press releases of one or two pages and short-format reports intended to make complex statistical data available to the general public have become popular during the last decade. The U.S. Bureau of the Census' *Census Briefs*, the National Science Foundation's *Issue Briefs*, and the National Center for Education Statistics' *Issue Brief* and *Statistics in Brief* series provide recent examples of short print report products aimed at a broad audience.

Descriptive analyses featuring tabular presentations are often released by statistical agencies. These vary in length, number, and complexity of the tabular information presented, but rarely provide complex statistical analyses. Typically, these reports describe results in narrative form, display both simple and complicated data tables, illustrate aspects of the data through graphical representation, or use combinations of these reporting formats.

Complex substantive data analyses, using techniques such as regression analysis and categorical data analysis, are often made available in what may be characterized as analytical reports. While these reports may also present descriptive data, their focus is answering specific questions and testing specific hypotheses. These reports and those mentioned above often rely on a single data set for the analysis, although this need not be the case. Compendium reports provide data from a large number of data sets, usually in the form of data tables and charts on a large and diverse set of topic areas.

Methodological/technical reports are released to describe special studies, and provide results of research on statistical and survey methods. These are reports released to provide specific information on a particular source of error or survey procedure, either to quantify the magnitude of the error source or to document a particular problem. Other technical reports do not focus on a specific methodological study, but rather provide substantial background information about data collection, processing, and estimation procedures. These reports are often characterized as "design and methodology" reports or user guides for the data collection program.

Although there is considerable recognition of the importance of reporting information about the survey design and the nature and magnitude of survey error sources, there is a lack of consensus about how much detail should be provided in the different reporting formats. Generally speaking, most survey methodologists and program managers agree that basic information about the survey's purpose, key variables, sample design, data collection methods, and estimation procedures ought to be available in descriptive and analytical reports. Additionally, there is a consensus that the sources of error described above, sampling error, nonresponse error, coverage error, measurement error, and processing error, should be described and accounted for when reporting results. There is less of a consensus and, perhaps no consensus, on the reporting of an error source when it is not a major source of error; for example, a discussion of coverage error in a survey where "state" is the sampling unit may not be necessary.

While there is general agreement that this information should be reported, there is no clear answer as to how much information to provide in the various reporting formats. A long, reasonably detailed discussion of error sources at the end of a lengthy complicated analytic report may seem reasonable in that context, but is obviously inappropriate for reports that may be only 2–10 pages long. Striking a reasonable reporting balance is the issue, because there is a strong belief that some information on the data source and error sources should be reported regardless of the length of the report. Obviously, details reported ought to depend on the nature of the report and its intended use.

An understanding of current practices in reporting the quality of survey data is largely dependent on anecdotal evidence and the experiences of individuals working within agencies and survey programs. The subcommittee addressed this limited understanding of current practices in reporting error sources by conducting three studies aimed at trying to characterize current agency practices in reporting sources of error. The results of these studies (Kasprzyk et al. 1999) provided a framework for a discussion of issues and recommendations. The studies dealt with three reporting formats: the short-format report, the analytic report, and the use of the Internet. In the sections that follow, the subcommittee's studies on the extent to which sources of error are reported in each of the three reporting formats are described and the results of the studies and the subcommittee's recommendations are presented. The recommendations concern the kinds of information on sources of error that should be included in each of these three reporting formats.

2.2 The Short-Format Report

2.2.1 The Short-Format Report Study

Short reports, directed to specific audiences, such as policymakers and the public, focus typically on a very narrow topic, issue, or statistic. McMillen and Brady (1999) reviewed 454 publications of 10 pages or less in length to examine their treatment of information about survey design, data quality, and measurement. The publications are products of the 12 statistical agencies that comprise the Interagency Council on Statistical Policy and are available over the Internet. The publications were released during the 1990s and were almost exclusively on-line reports. The majority of the reports were published in the mid-1990s or later. The reports reviewed from each agency were as follows: 91 from the Bureau of Labor Statistics; 79 from the Economic Research Service; 64 from the National Center for Education Statistics; 38 from the U.S. Bureau of the Census; 34 from the National Science Foundation; 28 from the National Agricultural Statistics Service; 24 from the Bureau of Justice Statistics; 22 from the Bureau of Economic Analysis; 22

from the Internal Revenue Service; 13 from the National Center for Health Statistics; 8 from the Energy Information Administration; and 3 from the Bureau of Transportation Statistics.

The study found considerable variation in the amount of documentation included across this reporting format. Since there is little consensus over the amount of detail to be included in short reports, the authors limited the review of the reports as to whether or not specific error sources or specific elements of documentation about survey design were mentioned. Virtually all of the short reports include some information on how to learn more about the data reported. This information ranged from name, phone number, e-mail and web site address to citations of printed reports.

Approximately two-thirds (69 percent) of the 454 reports included either a reference to a technical report or some mention of study design, data quality, or survey error. Close to one-half (47 percent) included some information describing the purpose of the survey or analysis, the key variables, the survey frame, and/or key aspects of the sample design. Only 20 percent included the sample size and 10 percent described the mode of data collection. Only a very small fraction (2 percent) mentioned estimation and/or weighting.

About one-fifth (22 percent) mentioned sampling error. In most cases, this was no more than a mention, although occasionally statistical significance testing and significance level were noted. Only a handful of reports included information on the size of the sampling error. Nonresponse error is the most visible and well-known source of nonsampling error and certainly the most recognizable indicator of data quality. Despite this, only 13 percent of the short reports included any reference to response rates, to nonresponse as a potential source of error, or to imputations. Only 3 percent reported unit nonresponse rates and there was virtually no reporting of item nonresponse rates. Coverage rates or coverage as a potential source of error was mentioned in only 10 percent of the reports covered. The difficulties associated with measurement were reported in 22 percent of the reports reviewed. Processing errors as a potential source of survey error were cited in 16 percent of the reports.

Results of the study are not surprising. The types of reports studied are short and oriented to a specific topic. The principal goal of the publication is to convey important policy-relevant results with a minimum of text. Discussion of sources of error in this report format is not viewed as critical. However, the disparity between stated policy and implemented policy concerning the reporting of sources of error is obvious.

2.2.2 Short-Format Reports: Discussion and Recommendations

The short-format report presents limitations on the amount of information that can be presented. Nevertheless, the subcommittee felt the essential principle of reporting information on the nature and magnitude of error must continue to be addressed. The subcommittee recommends that:

> All short-format reports provide basic information about the data set used in the analysis, known sources of error, related methodological reports, and a contact for further information.

The information presented must, of necessity, be brief, yet it must contain enough salient information that the reader can appreciate the limitations of the methodology and data. Thus, the report should include the name and year of the data collection program the analyses are based on and whether the data are based on a probability sample or census. It should also state that the data reported are subject to sampling error (if a sample survey) and nonsampling error. The total in-scope sample size and the overall unit response rate should be reported. Reports having statements describing findings should state whether statistical significance testing was used and reference the significance level. It should include a statement that sampling errors for estimates in the reports are available on request. When only a few estimates are displayed, presenting confidence intervals associated with the estimates may be appropriate. Estimates dependent on survey variables with high item nonresponse rates or having particularly difficult measurement properties should be identified. A reference to a source report that includes more detailed information about data collection and data quality should be cited along with the name of a contact person who can provide additional information or answer questions.

The information in the recommendation can be conveyed in a short paragraph at the conclusion of a short-format report. The subcommittee recommends that agencies adopt a reporting format that can be repeated with only minor modifications across their short-format reports. One example might look like this:

> Estimates in this report are based on a national probability sample of <Sample Size> drawn from the < Sampling Frame>. All estimates are subject to sampling error, as well as nonsampling error, such as measurement error, nonresponse error, data processing error, and coverage error. Quality control and editing procedures are used to reduce errors made by respondents, coders, and interviewers. Statistical adjustments have been made for unit nonresponse and questionnaire item nonresponse. All differences reported are statistically significant at the 0.05 level. The response rate for the survey was xx.x percent. Sampling errors for the estimates in this report are available from <Sampling Statistician (phone number; e-mail address)>. Detailed information concerning the data collection (including procedures taken to test the wording of questions), methodology, and data quality are available in <Data Collection and Methodology Report>. For more information about the data and the analysis contact <Program Contact (phone number; e-mail address)>.

2.3 The Analytic Report

2.3.1 The Analytic Report Study

A second study conducted by the FCSM subcommittee focused on a review of "analytic publications"—publications resulting from a primary summarization of a one-time survey or an ongoing series of surveys. Analytic publications may use a variety of formats with results described in narrative form, displayed in tables, shown in graphical format, or a combination of these. Atkinson, Schwanz, and Sieber (1999) conducted a review of 49 analytic publications produced by 17 agencies. The review included publications from major statistical agencies, as well as some from smaller agencies conducting surveys. The selected publications were a convenience sample, but an effort was made to cover as many of the major statistical agencies as time would allow.

The review considered both the completeness of background information on survey design and procedures and reports of error sources. Evaluation criteria were established for the kinds of survey information that ought to be included in analytic reports. The fifty-one review criteria identified are listed in tables 2.1 and 2.2. For each of these, a value of “1” or “0” was assigned to each criterion, depending on whether or not the publication contained the qualifying information for the criterion. To facilitate, standardize, and document the review of each publication, hierarchical levels of increasing detail about each of the major categories of error and background survey information were also established. The criteria for sources of error consisted of a three-level hierarchy ranging from level 1, where a particular error source was merely mentioned through level 3, where detailed information about the error source was provided (table 2.1). Levels 2 and 3 generally involved some quantification of the error source.

Sampling error was the most frequently documented error source, being mentioned in 92 percent of the reports. Among the analytic reports reviewed, 75 percent presented sampling errors, 75 percent gave a definition and interpretation, and 45 percent specified the method used in calculating sampling errors. Somewhat surprisingly, only 71 percent mentioned unit nonresponse, 59 percent reported an overall response rate, and 20 percent reported response rates for subgroups. Only one-half (49 percent) mentioned item nonresponse and only 22 percent reported any item response rates. Nearly all reports included a definition of the universe (94 percent) and identified and described the frame (84 percent), but only one-half (49 percent) specifically mentioned coverage error as a potential source of nonsampling error, and only 16 percent provided an estimated coverage rate. Two-thirds of the reports mentioned measurement error and one-half included a description and definition. Specific studies to quantify this error were mentioned in only 18 percent of the reports. The majority of the reports (78 percent) mentioned processing as an error source, but very few included any detail about this error source (about 4 percent reported coding error rates and 6 percent reported edit failure rates).

A second set of criteria was defined to measure the extent to which contextual survey information is included in publications. This information explains the survey procedures and helps the reader understand the survey results and their limitations. Two levels of increasing detail were defined for each of the four categories of background survey information (table 2.2).

The study indicated that survey background information was reported reasonably well—the general features of the sample design were reported about 92 percent of the time, data collection methods about 88 percent of the time, and a brief description of the estimation techniques about 82 percent of the time. The review of error sources revealed variation across agencies. Only 59 percent of the reports included at least some mention of each of the five error sources.

The results of this study are not comforting. While recognizing the subjective nature of the evaluation criteria and the obvious limitations of a small convenience sample, the fact remains that a considerable discrepancy exists between stated principles of practice and their implementation when it comes to the nature and extent of reporting sources of survey error.

2.3.2 The Analytic Report: Discussion and Recommendations

Analytic reports, as we have defined them for this study, include a wide variety of report series and types of analysis. The most important characteristic of this kind of report is that it provides fairly detailed analyses and/or summaries of data from either one-time or continuing surveys. The reports, themselves, are longer than the reports described in section 2.2 and provide more opportunity for data providers and analysts to describe the sources and limitations of the data. The subcommittee’s recommendations take advantage of this fact while recognizing that the

Table 2.1.—Evaluation criteria for the sources of error

Error type	Level	Criteria
Coverage error	1	Coverage error is specifically mentioned as a source of nonsampling error
	2	Overall coverage rate is provided Universe is defined Frame is identified and described
	3	Coverage rates for subpopulations are given Poststratification procedures and possible effects are described
Nonresponse error	1	Unit nonresponse is specifically mentioned Item nonresponse is specifically mentioned Overall response rate is given
	2	Item response rates are given Weighted and unweighted unit response rates at each interview level are given Numerator and denominator for unit and item response rate are defined
	3	Subgroup response rates are given Effect of nonresponse adjustment procedure is mentioned Imputation method is described Effect of item nonresponse is mentioned Results of special nonresponse studies are described
Processing error	1	Processing errors are specifically mentioned
	2	Data keying error rates are given Coding error rates are given Edit failure rates are summarized References are given to processing error studies and documentation
	3	Coder variance studies or other processing error studies are given
Measurement error	1	Measurement error is mentioned as a source of nonsampling error
	2	Specific sources of measurement error are described and defined
	3	Reinterview, record check, or split-sample measurement error studies are mentioned and/or summarized with references to larger reports
Sampling error	1	Sampling error is mentioned as a source of error Definition and interpretation of sampling error is included Significance level of statements is given
	2	Sampling errors are presented Confidence intervals are defined and method for calculating intervals is described Sampling errors and calculations for different types of estimates (e.g., levels, percent, ratios, means, and medians) are described
	3	Method used for calculating sampling error is mentioned with reference to a more detailed description Generalized model(s) and assumptions are described

Table 2.2.—Evaluation criteria for background survey information

Error type	Level	Criteria
Comparison to other data sources	1	General statement about comparability of the survey data over time is included General statement about comparability with other data sources is included Survey changes that affect comparisons are briefly described
	2	Survey changes that affect comparisons of the survey data over time are described in detail Tables, charts, or figures showing comparisons of the survey data over time are included Tables, charts, or figures showing comparisons with other data sources are included
Sample design	1	General features of sample design (e.g., sample size, number of PSUs and oversampled populations) are briefly described
	2	Sample design methodologies (e.g., PSU stratification variables and methodology, within PSU stratification and sampling methodology and oversampling methodology) are described in detail with references to more detailed documentation
Data collection methods	1	Data collection methods used (e.g., mail, telephone, personal visit) are briefly described
	2	Data collection methods are described in more detail Data collection steps taken to reduce nonresponse, undercoverage, or response variance/bias are described
Estimation	1	Estimation methods are described briefly
	2	Methods used for calculating each adjustment factor are described in some detail Variables used to define cells in each step are mentioned Cell collapsing criteria used in each step are mentioned

purpose of the report is to present statistical information. The length limitations found in the short-format reports do not apply here and, consequently, an opportunity exists for a fuller treatment of descriptions of the survey, methodology, and data limitations. On the other hand, the fuller treatment of methodology in an analytic report cannot be so lengthy and detailed that this information overshadows the statistical information presented in the report.

Analytic reports are usually intended for a broad and multi-discipline audience. The reports usually provide a technical notes or methodology appendix containing information about the data sources and their limitations. A critical aspect of the recommendations is the understanding that information presented in a technical or methodology appendix must provide the essentials or key aspects of the survey background and the major sources of error in the survey. The details of the data collection operations and procedures, studies about the error sources, and detailed analyses of the effects of statistical and procedural decisions belong in individual technical reports, comprehensive design and methodology reports, or quality profiles. The technical appendix does not need to be lengthy, 5–10 pages, but it should provide quantitative information to inform the reader as well as citations to secondary sources that provide more detailed information or analyses.

The subcommittee recommends that studies reporting analyses of statistical data should present three types of information:

- Background description of the data collection programs used in the analysis (table 2.3 lists key information),
- Description of each major source of error, the magnitude of the error source (if available), and any known limitations of the data (table 2.4 lists essential information), and
- Access to the questionnaire or questionnaire items used in the analysis, through the report or through electronic means, or upon request.

The information described in the recommendation above helps readers/users of the report to better understand the report's findings. The subcommittee appreciates that a substantial amount of material is typically available on these topics in the data collection specifications. The difficult task for the data producer is to synthesize the available material into a short technical appendix.

Table 2.3.—Background survey information

Survey objectives	
Survey content	
Changes in content, procedures, and design from previous rounds	
Survey preparations/pretests	
Sample design	
	Target population defined
	Sampling frame identified and described
	Stratification variables
	Sample size
Data collection	
	Schedule (when collected/number of follow-ups/time in field)
	Mode (percent of each type)
	Respondent (identified/percent self/ percent proxy)
	Reference period identified
	Interview Length
Data processing	
	Identification of procedures used to minimize processing errors
	Editing operations
	Coding operations
	Imputation methods
Estimation	
	Description of procedure (stages of estimation)
	Source and use of independent controls
Key variables/concerns defined	

Table 2.4.—Limitations of the data

- Sampling error
  - Described
  - Interpreted
  - Calculation method stated
  - Presentation of sampling error
    - Sampling error tables
    - Generalized variance model parameters
    - Design effects
  - Description of how to calculate sampling error

- Nonsampling error
  - Description
  - Sources identified

- Nonresponse error
  - Definition of total unit response
    - Numerator and denominator specified
    - Assumptions for the calculation of the response rate stated (RDD, for example)
    - Special situations clarified (defining longitudinal response rates, for example)
  - Unit response rates (unweighted and weighted) at each level reported
  - Overall response rate reported
  - Special nonresponse studies cited if low unit response rates occur
  - Item response rates summarized

- Coverage error
  - Coverage error defined
  - Target population defined
  - Study population defined
  - Sampling frame identified (name and year)
  - Coverage rates (population/subpopulation rates) provided

- Measurement error (summarize results and refer to technical reports)
  - Measurement error defined and described
  - Special studies identified (reinterview studies/record check studies)
  - Technical reports/memoranda referenced

- Processing error
  - Processing error described
  - Data entry (keying/scanning) error rates
  - Edit failure rates (summarized)
  - Coding error rates

- Comparison to other data sources
  - Identification/description of independent sources for comparisons
  - Tables/charts/figures comparing estimates
  - Limitations of comparison tables described

- References about the data collection program, survey and sample design, error sources, and special studies

The second type of information that ought to be included in a technical appendix concerns information about the accuracy of the estimates presented in the report. All statistical agencies address this issue in some fashion. However, the information is presented inconsistently. Basic statistical data to inform users of the quality of the data collection operations are often not reported. Substantial gaps exist in the reporting of quantitative information. Three aspects of the estimates should be addressed in the technical appendix: sampling error, nonsampling error, and comparisons with other data sources.

Sampling error should be defined and presented. Access to sampling errors of the survey estimates should be provided. Nonsampling error should be described and, since in most cases it is difficult to quantify, statistical indicators that serve as proxies for its actual measurement should be presented. A short discussion of each error type presented in the report along with any available data about the extent of that error type should be summarized for the data user. Presenting statistical information about the error source is important—either direct information about the error source or proxy information. Table 2.4 identifies some important topics that ought to be included in a discussion of sources of error. The list is lengthy, but a detailed treatment of each topic is not what is being advocated. Finally, if comparable data are available, information about the comparisons should be provided and detailed analyses referenced.

The third piece of information that should be made available in the appendix is the questionnaire itself, the questionnaire items used in the analysis, or at a minimum access to questionnaires, perhaps electronically or upon request. The availability of the questionnaire allows the reader to understand the context of the question asked of the respondent.

2.4 The Internet

2.4.1 The Internet Study

The Internet has become the principal medium for the dissemination of data products for most federal statistical agencies. The third study (Giesbrecht et al. 1999) reviewed guidelines and practices for reporting error sources over the Internet. Some federal agencies have written standards for Web sites, but these generally focus on Web site design, layout, and administrative access. A few agencies, such as the U.S. Bureau of the Census, have begun the process of developing standards for providing information about data quality over the Internet (U.S. Bureau of the Census 1997). This draft report gives details of data quality information that ought to be provided to the user, but does not require or suggest the use of Internet features for making information more accessible. Generally, standards documents related to Internet practices reiterate standards for printed documents (for example, United Nations Economic and Social Council 1998).

The study reviewed the accessibility of data quality documentation on current Internet sites of 14 federal agencies with survey data collections. Online data documentation was available for most of the sites visited (78 percent). For about one-half the sites, offline documentation was referenced as well. Most agencies seem to upload their printed reports and documentation in the form of simple text or Adobe Acrobat portable document format (PDF) files. In addition, one-half the sites offered technical support online and an additional 29 percent included lists of telephone contacts on their web sites. The study also noted a few best practices found on the visited Web sites, such as the availability of pop-up windows providing definitions of column and row headings in tables, links to send e-mail messages to technical specialists, links to "survey methodology" and "survey design" documentation, explicit directions to users about errors and comparability issues, links from one agency's home page to another, and common access points to statistical information.

The study found current Internet standards for data quality information echo the standards for printed reports and statistical tables. More explicit guidelines for how the advantages of the Internet medium should be employed to make data quality information more accessible do not seem to exist. The development of metadata standards (Dippo 1997), however, as an integral part of the survey measurement process may facilitate the creative use of the Internet.

2.4.2 The Internet Study: Discussion and Recommendations

The use of the Internet for reporting statistical information is growing and evolving so fast that recommendations seem inappropriate since they become out-of-date very quickly. The potential use of this new medium has not been fully developed, and statistical agencies while providing much information on the Internet have only begun to explore its potential. In general, agencies report electronically what is reported on paper, often in the form of PDF files that are no more interactive than the paper report. Thus, the limitations on reporting information on the data collection program and sources of error are limitations of the printed report itself.

Large gaps exist between the potential of the medium and implementation within the medium. The key issue is how to organize and display statistical information and its corresponding documentation in a way that can be understood and easily accessed by the user community. Thus, it is important for statistical agencies to continue developing online design features, such as frames, audio/video, hyperlinks to relevant documents (such as design, estimation, and technical documentation) or parts of the same document, pop-up windows (for, among other applications, providing data definitions for terms in tables or for providing the sampling error of the estimate in the table), online data analysis tools, user forums, and e-mail technical support links to improve service to data users.

The subcommittee recommends:

> Agencies should systematically and regularly review, improve access to, and update reports and data products available on the Internet, particularly to reports about the quality of the data; the amount of information about data quality should be no less than that contained in printed reports; linkage features available on the Internet, such as hypertext links, should be used to improve access to information about the data collection program and its sources of error. Information displayed on the Internet should incorporate good design principles and "best practices" of displaying data and graphics on the web.

Agency practices will dictate whether the Internet reporting function is decentralized or not. Either way, financial and staff resources should be allocated to developing new applications to improve online access to information about the quality of data in reports and products on the Internet.

Predicting future development is difficult, however, as printing and traditional dissemination costs continue to increase and Internet access in households continues to grow. In fact, it may become increasingly common to find that information is available *only* through the Internet. Internet dissemination ought to spur the development of new ways to present statistical information and new ways to inform data users about the quality of the statistical information. At this time, based on our review of Internet sites, the paper report model is almost universal. The Internet product is developed after the paper product is completed. This suggests to us that the potential of the Internet to present and display information has not been addressed from the point of view of a dissemination plan based solely on the Internet. Otherwise, the use of video, audio, frames, and hyperlinks would be more obvious. Consequently, we suggest that data, reports and press releases available only though the Internet be developed to take maximum advantage of the new medium.

2.5 General Observations

Specific recommendations about reporting the nature and extent of sources of error in data collection programs are highly dependent on the form of the report. Reporting limitations are apparent in light of the different formats for releasing statistical information. The fundamental issue is the identification of critical information about the data collection program and the principal sources of error likely to affect an analyst's appreciation of the results. The development of general recommendations reported in this chapter highlights the practical difficulties of reporting about data collection programs and their error sources when reporting formats place limitations on the amount of information reported. The remainder of the report will no longer address the short-format report, but will focus on the analytic or substantive report.

The analytic report in its many variations provides broad coverage of a substantial number of analyses and topics. The length and format of these reports provide the survey methodologist and survey statistician an opportunity to inform data users about the quality of data. Recommendations concerning analytic reports provide the minimum amount of information the subcommittee thought ought to be available. In the course of specifying minimum requirements, the subcommittee recognized that much more information about the survey program and its sources of error should be available to the data user. Thus, each chapter has two sets of recommendations: 1) the minimum reporting requirements about the survey program and its error sources for analytic reports; and 2) full reporting requirements about error sources in the context of more comprehensive documents such as methodological reports, user manuals, and quality profiles.

References

Atkinson, D., Schwanz, D., and Sieber, W. K. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Dippo, C. 1997. "Survey Measurement and Process Improvement." In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality.* New York: John Wiley & Sons. 457–474.

Giesbrecht, L., Miller, R., Moriarity, C., and Ware-Martin, A. 1999. "Reporting Data Quality on the Internet." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 342–354.

Kasprzyk, D., McMillen, M., Atkinson, D., Giesbrecht, L., Schwanz, D., and Sieber, W.K. 1999. "Reporting Sources of Error: The United States Experience." *Proceedings of the 52nd Session of the International Statistical Institute.* International Association of Survey Statisticians. 19–30.

McMillen, M. and Brady, S. 1999. "Reporting Sources of Error in Short Format Publications." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 316–328.

United Nations Economic and Social Council. May 18–20, 1998. "Guidelines for Statistical Metadata on the Internet." Paper contributed to the Conference of European Statisticians, Forty-sixth Plenary Session. Paris.

U.S. Bureau of the Census. 1997. *Survey Design and Statistical Methodology Metadata IT Standards* (Draft). Washington, DC: U.S. Department of Commerce.

Chapter 3

Sampling Error

3.1 Introduction

Sampling error refers to the variability that occurs by chance because a sample is surveyed rather than all the units in the population. Sampling error is probably the best-known source of survey error as evidenced by its recognition by the popular press.

In particular, sampling error refers to the expected variation in estimates due to the random selection scheme used to select the sample. In a random selection scheme, each unit of the population has a known, non-zero probability of being selected into the sample. The method of randomization is important because it can be used in theory to define both the optimal estimator and the appropriate estimate for sampling error. Most federal surveys using random selection are designed so that sampling errors can be computed directly from the survey observations. In other situations, the design or estimate is so complex that approximation methods must be used to estimate sampling errors.

While most federal surveys and virtually all demographic surveys use specialized random sample selection mechanisms, there are establishment surveys that make use of "cut-off" samples—a sample consisting of only the largest establishments. For these surveys, a traditional sampling error estimate is not defined because randomization was not used in selecting the sample. Typically, cut-off samples are used to estimate the aggregate of items such as sales, deliveries, or revenues from populations that are highly skewed. If the quantity covered by the sample is sufficiently high (80 percent or so) and data from the recent past for the population units are available, then cut-off samples along with ratio estimation may perform quite well. For such aggregates, "estimation error" can be computed from a regression model relating the values reported by companies at different points in time. In these situations, a model-based estimate of error may be used to describe the accuracy of the estimated totals.

Reporting the existence and magnitude of sampling errors or "estimation error" along with the estimates allows users of surveys conducted by the federal government to make more informed policy decisions. The regular preparation and presentation of measures of the precision of estimates from federal surveys also support other goals, such as the evaluation and improvement of the survey design.

Section 3.2 of this chapter addresses methods of estimating sampling error. Approaches for presenting sampling error are discussed in section 3.3. Section 3.4 addresses the practices of federal agencies in the presentation of sampling error estimates.

3.2 Measuring Sampling Error

The sampling error for estimates produced from simple random samples can be computed easily. However, federal surveys are usually not based on simple random samples because of cost constraints and the requirement to produce reliable estimates for subgroups. For example, the U.S. Bureau of Labor Statistics' and the U.S. Bureau of the Census' Current Population Survey and the U.S. Bureau of Labor Statistics' Consumer Expenditure Survey involve personal interviews (U.S. Bureau of Labor Statistics 1997). It would be costly to hire and train staff to interview samples that are widely dispersed across the country, as would be the case with simple random samples. Instead, multistage sample designs are used to cluster the samples so that field work can be accomplished economically. The methods for computing sampling errors must account for this clustering.

Federal surveys also often have multiple objectives including producing precise estimates for subgroups of the population. With a simple random sample, the size of the total sample has to be increased to accomplish this goal and that drastically increases the cost of the survey. The alternative is to sample certain segments of the population at different rates using unequal probability sampling schemes. For example, the National Health Interview Survey uses higher sampling rates for blacks and Hispanics to improve the precision of the estimates for these subgroups without increasing the overall sample size (National Center for Health Statistics 2000). These types of sampling schemes affect the estimation of sampling errors.

Establishment surveys conducted by the federal government also do not use simple random samples because the population of establishments is skewed. That is, most businesses are small (in terms of sales volume, number of employees, etc.). A simple random sample of businesses likely would consist almost entirely of small businesses. Hence, estimates of totals (sales volume, number of employees, etc.) from a simple random sample would not be as precise as estimates using other sampling procedures—for example, sampling probability proportional to size (Cochran 1977) or stratified sampling, where large businesses can be sampled from a "large business" stratum (see, for example, Hansen, Hurwitz, and Madow (1953), volume 1, chapter 12).

Both complex estimates from traditional sample designs and simple estimates from complex sample designs, may require special approaches to variance estimation. A brief overview of the methods used is given below. Those interested in greater detail can consult standard references such as Cochran (1977); Hansen, Hurwitz, and Madow (1953); Kish (1965); Sarndal, Swensson, and Wretman (1992); and Wolter (1985).

3.2.1 Variance Estimation Methods

Most federal surveys are designed so that the key statistics can be precisely estimated from the sample and the sampling error of those estimates can also be computed from the survey itself. This implies that the sample sizes are large enough that the estimates satisfy the requirements of the large-sample statistical theory developed for sample surveys. For very small subgroups with small sample sizes, this approach to estimating the variance or sampling error of the estimates may not be appropriate. Sarndal, Swensson, and Wretman (1992) discuss these issues more completely. In this chapter, relatively large sample sizes are assumed.

There are two main methods of estimating sampling errors when estimates based on sample surveys are too complex to support direct estimation of variances: the Taylor series linearization and replication. The methods are both approximations in practice, with different benefits and drawbacks depending on the sample design and statistic being computed.

Taylor series linearization has been used to produce variance estimates for many federal surveys. For example, this methodology was used to produce variance estimates for the dual system estimates computed from data from the 1988 Post-Enumeration Surveys conducted at the U.S.

Bureau of the Census (Moriarity and Ellis 1990). Nonlinear statistics, such as ratios of estimates, and even many means and percentages estimated for subgroups, are commonly estimated in federal surveys. The Taylor series linearization procedure simplifies the variance estimation problem by replacing the nonlinear statistic by its first-order (linear) Taylor series approximation. The variance of the linear approximation is computed using standard methods for the sample design. More detail on the theory and practice of using Taylor Series linearization can be found in Wolter (1985).

The second method of variance estimation for complex surveys is replication. In replication, the sample is partitioned into subsamples or replicates and the statistic of interest is computed for each of these replicate samples. The variation between the estimates from the replicates is used to estimate the variance of the estimate computed from the full sample. Replication methods generally take account of sample design features such as stratification and primary sampling unit (PSU). Different replication methods exist: balanced repeated replication, the jackknife, random groups, and the bootstrap. The bootstrap method, a computer-intensive method for estimating or approximating the sampling distribution of a statistic and its characteristics, can be used in applications to complex survey data, and is described in Rao and Wu (1988). Wolter (1985) provides descriptions of the other methods.

Replication methods have been used to produce variance estimates for many federal government surveys. For example, this method has been used for many years to produce variance estimates from the Current Population Survey, as described in U.S. Bureau of Census and U.S. Bureau of Labor Statistics (2000). Many surveys at the National Center for Education Statistics use replication methods. See Gruber et al. (1996) for the use of the method in the 1993–94 Schools and Staffing Survey.

3.2.2 Computer Software for Variance Estimation

In recent years, software for computing sampling errors from complex surveys such as those conducted by the federal government have become available for wide-spread use. These software packages implement either the Taylor series or replication methods of variance estimation and require the user to identify essential design variables such as strata, clusters, and weights. The information required to use the software depends on the method of variance estimation and the way it is handled in the specific package. For the Taylor series method most software products require the survey microdata file to contain the sample weight, the variance stratum, and the PSU. Software for replication methods requires either the same information or replicate weights.

A World Wide Web site that reviews computer software for variance estimation from complex surveys was created with the encouragement of the Section on Survey Research Methods of the American Statistical Association {http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html}. The site lists software packages that are available for personal computers and provides direct links to the home pages of these packages. The site also contains articles that provide general information about variance estimation and articles that compare features of the software packages.

3.2.3 Sampling Error Estimates from Public Use Files

Many federal agencies produce public use files that enable analysts to obtain subsets of the survey data and conduct their own analyses. The public use files often contain the information needed to compute sampling errors using Taylor series or replication methods. The documentation accompanying the public use files describes how the data on the public use file can be used to compute sampling errors. For example, the National Household Education Survey (NHES), conducted by the National Center for Education Statistics (NCES), provides documentation and replicate weights so that the sampling errors can be estimated from the public-use files (U.S. Department of Education 1994). In this case, the NHES defined 60 replicates of the sample based on the sample design of the survey and created 60 replicate weights using the same estimation procedures as the full sample. These replicate weights are included on the data files, and computation of individual sampling errors is possible with variance estimation software. Replicate weights (100 replicates) with instructions for their use in sampling error estimation are also available on the public use data file for the National Survey of Family Growth, Cycle IV (National Center for Health Statistics 1990).

Sometimes the design information needed for variance estimation might not be included on the public use file because of concerns that it might be used to identify some of the respondents to the survey. For example, the variance stratum and PSU identifiers along with survey microdata might let a knowledgeable user identify respondents. One solution to this problem is to modify the design information in some way before including it on the public use file so that such disclosure is not possible. This is done by collapsing or blurring the variance stratum or PSU and including the modified versions of the data on the public use data files. This allows the computation of reasonably accurate variance estimates, but no identifying information is released. For example, the National Center for Health Statistics (NCHS) produced a public use file for the 1995 National Health Interview Survey with collapsed variance stratum and PSU. The file documentation includes an article by Parsons (1998) that describes how to use this information and notes that the sampling errors from the public use file will not agree exactly with those published using the file without the collapsing. Similarly, the user's guide for the Survey of Income and Program Participation (U.S. Bureau of the Census forthcoming) describes variance units and variance strata that can be used for variance estimation purposes.

Other public use files from federal surveys do not support user computation of sampling errors. Several alternatives exist for this situation. One alternative is to supply generalized variance functions (discussed in section 3.3.2). For example, the National Center for Health Statistics' National Survey of Family Growth, Cycle IV, provides generalized variance functions in the survey's public use data tape documentation (National Center for Health Statistics 1990). Another alternative is to supply "design effects." A design effect is defined as the ratio of the design-based sampling variance to the sampling variance under simple random sampling, assuming the same sample size. A public use data file user can calculate sampling variances assuming simple random sampling and then multiply by the appropriate design effect to obtain a valid estimate of a design-based sampling variance (see, for example, Salvucci and Weng 1995; Salvucci, Holt, and Moonesinghe 1995; Ingels et al. 1994). A third alternative is to supply bootstrap samples to users. Bootstrap weights are constructed using "internal" information (see the 1993 National Survey of Small Business Finances on the Internet at http://www.bog.frb.fed.us/pubs/oss/oss3/nssbf93/bootsrp.html).

3.3 Approaches to Reporting Sampling Error

As noted in the introduction, reporting on the nature and extent of sampling errors is an essential part of the presentation of the survey findings. Gonzalez et al. (1975) is an excellent reference concerning reporting errors in analytic publications; they take a comprehensive approach by including both sampling and nonsampling errors.

All publications and other releases of data such as public use files from sample surveys sponsored by the federal government should have appropriate statements to inform users that the estimates are subject to sampling error. The reports and documentation should define and interpret all terms such as sampling errors so that users grasp that the estimates are not exactly equal to the population quantities being estimated. This information is needed because such errors might affect the conclusions drawn from the survey.

Statements of sampling error that are commonly used in reports from federal surveys indicate that the observed sample is just one of a large number of samples of the same size that could have been selected and that estimates from each sample would differ from each other by chance. For example, in a typical U.S. Bureau of the Census analytic publication (for example, Fronczek and Savage 1991; Norton and Miller 1992; Lamison-White 1997), the “Source and Accuracy” appendix of the report usually has a statement similar to the following from Gonzalez et al. (1975):

> The particular sample used in this survey is one of a large number of all possible samples of the same size that could have been selected using the sample design. Estimates derived from the different samples would differ from each other. The difference between a sample estimate and the average of all possible samples is called the sampling deviation. The standard or sampling error of a survey estimate is a measure of the variation among the estimates from all possible samples, and thus is a measure of the precision with which an estimate from a particular sample approximates the average result of all possible samples.
>
> The sample estimate and an estimate of its standard error permit us to construct interval estimates with prescribed confidence that the interval includes the average result of all possible samples (of a given sampling rate).

Sampling errors are often easier to interpret when the estimates are presented along with confidence intervals. For example, reports that give 95 percent confidence intervals should state that approximately 19/20 of the intervals constructed from all possible samples would include the average value over all possible samples. Confidence intervals provide a concise and effective way of communicating about errors in the estimates due to sampling. Unqualified point estimates in reports, on the other hand, imply a false degree of exactitude in the estimates.

Gonzalez et al. (1975) provide several examples of ways of presenting and interpreting sampling errors from federal surveys. The style of presentation may vary. Sometimes estimates of sampling error, or multiples of the sampling error, are reported in tables or graphs. Other times confidence intervals or graphical displays of confidence intervals are provided.

Despite this variation in style, the two methods of presenting information on the precision of the estimates involve either the reporting of a direct estimate of the error or an indirect estimate of the error reporting of generalized variance functions or average design effects that permit users to compute the errors for the estimates for each estimate.

3.3.1 Direct Estimates of Sampling Errors

The direct method of presentation involves computing an estimate of the sampling error for every statistic in a report using the techniques discussed above. For example, a public use CD-ROM product from the 1997 American Community Survey conducted by the U.S. Bureau of the Census provides information on sampling errors for every estimate provided on the CD-ROM {http://www/census.gov/acs/www/html/dataprod/1997}. Similarly, the NCES often provides corresponding standard error tables for every table of estimates in a report (Bobbitt, Broughman, and Gruber 1995).

This approach enables readers to view both the estimate and its sampling error at the same time so the variability in the estimates is clearly demonstrated. Of course, this makes the printed publication larger to accommodate the presentation of direct estimates of sampling variables with every point estimate. It is also worth noting that the sampling errors themselves are subject to sampling variation and may not be very stable, especially when effective sample sizes are small.

3.3.2 Indirect Estimates of Sampling Errors

Another presentation method used in some reports from federal surveys is to include only the estimates in the report and provide a procedure for the reader to compute approximate sampling errors for the estimates presented. This method is attractive because it saves space in the publication that otherwise would be required for the sampling errors. It may also allow users to estimate sampling errors for estimates that are not specifically reported in the publication but are simple functions of estimates that are given.

Two disadvantages of this approach are that the sampling errors are not presented physically close to the estimates and readers, thus, may not appreciate the level of uncertainty associated with the estimates. The other disadvantage is that the procedure used to compute the approximate sampling errors may not give values that are as accurate as direct estimates. See, for example, Bye and Gallicchio (1988).

One of the methods of presenting indirect estimates of sampling errors uses generalized variance functions (GVFs), model-based estimators of sampling variability. That is, given a data set of direct estimates of sampling variability, a model is fit to the data and then the model is used to make estimates of sampling variability. Typically, groups of estimates are formed and a separate model is fit for each group. Generalized variance functions have been popular because they supply a mechanism for computing large numbers of sampling error estimates rather easily, and require a minimum amount of space for presentation and explanation.

The National Center for Health Statistics, for example, includes a "technical notes on methods" appendix in its analytic reports. This appendix includes generalized variance functions along with instructions for their use. In the National Health Interview Survey (National Center for Health Statistics 1995), standard errors were computed for a broad spectrum of estimates. Regression methods were then applied to produce generalized variance equations from which a standard error for any estimate can be approximated. For a given estimate of characteristic *x,* associated model parameters *a* and *b* and a generalized variance function are given in the appendix. Approximate sampling errors of the given characteristic can then be computed using the formula $SE(x)=sqrt(ax^2+bx)$. Another discussion on the use of generalized variance functions can be found in the design and methodology report for the Current Population Survey (U.S.

Bureau of the Census and U.S. Bureau of Labor Statistics 2000). Generalized variance functions have shown in some data settings to perform as well or better than direct variance estimators in terms of bias, precision, and confidence interval construction (Valliant 1987).

The second indirect method uses average design effects. The design effect is the ratio of the sampling variance computed taking account of the complex design and estimation procedures to the sampling variance computed assuming the same size sample was selected as a simple random sample. Kish (1965) suggests that many estimates from a survey should have approximately the same design effect. An average design effect can be computed and then used to estimate the sampling error of an estimate in a report. Like GVFs, the average design effects are typically computed separately for different subgroups to improve the precision of the estimate. The *1993 National Household Education Survey School Readiness Data File Users' Manual* (U.S. Department of Education 1994) discusses the use of design effects as approximate sampling errors. The manual proposes several design effects for use depending on the nature of the estimate, where all design effects range between 1.0 and 1.5. Design effect tables are also provided for a number of population subgroups in the Schools and Staffing Survey (Salvucci and Weng 1995; Salvucci, Holt, and Moonesinghe 1995; Ingels et al. 1994). Wolter (1985) describes the theory and methods of constructing GVFs and average design effects.

A third method for providing indirect estimates of sampling error is to display the standard errors for a selected number of key estimates (Dalaker and Naifeh 1998). This approach has the advantage of an economic display of standard error information, but has the disadvantage that the user must extrapolate estimates of standard errors for estimates in which the standard errors are not provided.

3.4 Reporting Sampling Error in Federal Surveys

Sampling error is probably the best known source of survey error. The reporting of sampling error for survey estimates is important to all statistical agencies. For any survey based on a probability sample, data from the survey can be used to estimate standard errors of a statistic. At this point in time, software that takes account of the sample design is widely available to compute standard errors. There is substantial recognition of this source of survey error, evidenced by the fact that it is now recognized by the popular press. Linked to the calculation and reporting of sampling error is the issue of statistical significance testing in reports produced by federal agencies. Testing for statistical differences between survey estimates, differences over time and between subgroups, for example, is a critical feature of all survey programs and an important aspect of the sampling error reporting issue.

In the subcommittee's study of selected analytic publications, Atkinson, Schwanz, and Sieber (1999) found that sampling error was the error source most often discussed in these publications. Analytic reports provide more complete reporting of substantive results, either through text and/or tables, and usually have no page limitation. Hence, they have fewer constraints that may limit the discussion of survey errors.

Sampling error was discussed to some extent in 92 percent of the publications reviewed and presented in 74 percent of the publications. Thirty-seven publications (76 percent) provided the definition and interpretation of sampling error, while only 22 (45 percent) specified the method used for calculating the sampling errors. The form of analytic reports varies substantially from agency to agency. Constraints often exist between the availability of information on sources and magnitude of error and the timeliness and user friendliness of the report released to the public. Nevertheless, the results of the study of analytic reports are considered disappointing. The existence of such error should be acknowledged routinely in all publications and the presentation of sampling error should not be a matter of whether to present, but rather how to present such information.

The principal difficulty with the computation and presentation of sampling error occurs because of the multipurpose nature of many federal surveys. The computation and presentation of sampling errors for all survey estimates and differences between estimates is a major undertaking. In some publications, sampling errors are communicated in different ways, for example, as sample sizes, confidence intervals, and coefficients of variation. These are explained in the technical notes, survey notes, reliability statement, or appendices in the back of a report. Similarly, generalized variance functions (Dalaker and Naifeh 1998) and design effects are described in the technical notes of reports with instructions for their use as well. Some reports contain appendices of standard error tables that correspond to each table in a report (Bobbitt, Broughman, and Gruber 1995), while others present a standard error or confidence interval in the same table the estimate is presented (West, Wright, and Germino-Hausken 1995).

The availability of software to calculate direct estimates of variance has affected the reporting of information about sampling error. User's guides that advise data users analyzing public use files on methods for estimating standard errors have become important in recent years (U.S. Bureau of the Census forthcoming). These publications may include both methods for computing standard errors directly and methods for approximating the standard errors using other means, such as generalized variance functions. Electronic publishing holds great promise for new presentation formats for survey estimates and their sampling errors by displaying confidence intervals, sample sizes, or coefficients of variation through a simple click of a mouse.

The subcommittee recognizes the wide variation in users' requirements with respect to this source of error in surveys. Some users want to compute standard errors from design information provided on the public use data file, while others pay hardly any attention to the fact they exist. The correct balance of information provided on this subject is not easily determined. For analytic reports, however, the subcommittee recommends the following information be reported:

- Users must know if data are from a random sample or non-random sample; if the latter, then the implications for inference should be described.

- Sampling error should be identified as a source of error; it should be explained and interpreted for data users.

- If statistical tests are used in the report, the significance level at which statistical tests are conducted should be stated explicitly.

- Sampling errors for the principal estimates in a report should always be available to the reader; thus, tables of sampling errors, design effects, or generalized variance functions should be readily accessible, either through a presentation in the printed report or available electronically on the Internet.

- When space limitations preclude publishing detailed information, relevant technical publications should be provided as references, both print and electronic (URL) references.

Reporting more details concerning the development of sampling errors can be left to technical reports or user manuals. Consequently, the subcommittee recommends the following for inclusion in technical reports or user's manuals:

- The method used for calculating sampling error should be identified with reference to a more detailed description. If generalized models are used to provide sampling errors, the models, the assumptions underlying the models, and references to the results of the modeling should be available to the user of the data.

- Sampling error calculations for different types of estimates (e.g., levels, percents, ratios, means, and medians) should be described.

- Evaluations of the procedures used to estimate sampling errors should be described and discussed.

References

Atkinson, D., Schwanz, D., and Sieber, W.K. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Bobbitt, S., Broughman, S., and Gruber, K. 1995. *Schools and Staffing in the United States: Selected Data for Public and Private Schools, 1993–94.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 95–191).

Bye, B.V., and Gallicchio, S.J. 1988. "A Note on Sampling Variance Estimates for Social Security Program Participants from the Survey of Income and Program Participation." *Social Security Bulletin.* 51(10): 4–21.

Cochran, W. 1977. *Sampling Techniques*. New York: John Wiley & Sons.

Dalaker, J. and Naifeh, M. 1998. *Poverty in the United States: 1997.* Current Population Reports, Series P-60-201. Washington, DC: U.S. Bureau of the Census.

Fronczek, P. and Savage, H. 1991. *Who Can Afford to Buy a House.* Current Housing Reports*,* Series H121/91-1. Washington, DC: U.S. Bureau of the Census.

Gonzalez, M.E., Ogus, J. L., Shapiro, G., and Tepping, B.J. 1975. "Standards for Discussion and Presentation of Errors in Survey and Census Data." *Journal of the American Statistical Association.* 70 (351): Part II.

Gruber, K., Rohr, C., and Fondelier, S. 1996. *1993–94 Schools and Staffing Survey: Data File User's Manual, Volume 1: Survey Documentation.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 96–142).

Hansen, M., Hurwitz, W., and Madow, W. 1953. *Sample Survey and Methods and Theory*, *Vols. 1 and 2*. New York: John Wiley & Sons.

Ingels, S.J., Dowd, K.L., Baldridge, J.D., Stipe, J.L., Bartot, V.H., and Frankel, M. 1994. *National Education Longitudinal Study of 1988: Second Followup: Student Component Data File User's Manual*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 94–374).

Kish, L. 1965. *Survey Sampling.* New York: John Wiley & Sons.

Lamison-White, L. 1997. *Poverty in the United States: 1996.* Current Population Reports, Series P-60-198. Washington, DC: U.S. Bureau of the Census.

Moriarity, C. and Ellis, Y. 1990. "Documentation of the 1988 Dress Rehearsal Post-Enumeration Survey Estimation Process." STSD 1988 Dress Rehearsal Memorandum #V-20. Washington, DC: U.S. Bureau of the Census.

National Center for Health Statistics. 2000. "Design and Estimation for the National Health Interview Survey: 1995–2004." *Vital and Health Statistics.* Washington, DC: Public Health Service. 2(130).

National Center for Health Statistics. 1995. “Current Estimates from the National Health Interview Survey, 1994.” *Vital and Health Statistics*. Washington, DC: Public Health Service. 10(193).

National Center for Health Statistics. 1990. *Public Use Data Tape Documentation—National Survey of Family Growth, Cycle IV, 1988.* Washington, DC: Public Health Service.

Norton, A. and Miller, L. 1992. *Marriage, Divorce, and Remarriage in the 1990's.* Current Population Reports, Series P-23-180. Washington, DC: U.S. Bureau of the Census.

Parsons, V.L. 1998. “Variance Estimation for Person Data Using the NHIS Public Use Person Data Tape, 1995.” Available at http://www.cdc.gov/nchs/data/pvar.pdf

Rao, J.N.K. and Wu, C.F.J. 1988. “Resampling Inference with Complex Survey Data.” *Journal of the American Statistical Association.* 83(401): 231–241.

Salvucci, S., Holt, A., and Moonesinghe, R. 1995. *Design Effects and Generalized Variance Functions for the 1990–91 Schools and Staffing Survey, Volume II, Technical Report.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 95–342–II).

Salvucci, S. and Weng, S. 1995. *Design Effects and Generalized Variance Functions for the 1990–91 Schools and Staffing Survey, Volume I, User's Manual.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 95–342–I).

Sarndal, C., Swensson, B, and Wretman, J. 1992. *Model Assisted Survey Sampling.* New York: Springer-Verlag.

U.S. Bureau of Labor Statistics. 1997. *BLS Handbook of Methods*. Bulletin 2490. Washington, DC.

U.S. Bureau of the Census. Forthcoming. *Survey of Income and Program Participation User's Manual*. Washington, DC.

U.S. Bureau of the Census and U.S. Bureau of Labor Statistics. 2000. *The Current Population Survey: Design and Methodology.* Technical Paper 63. Washington, DC.

U.S. Department of Education. 1994. *National Household Education Survey of 1993: Early Childhood School Readiness Data File–User's Manual*. Washington, DC: National Center for Education Statistics (NCES 94–193).

Valliant, R. 1987. “Generalized Variance Functions in Stratified Two-Stage Sampling.” *Journal of the American Statistical Association.* 82: 499–508.

West, J., Wright, D., and Germino-Hausken, E. 1995. *Child Care and Early Education Program Participation of Infants, Toddlers, and Preschoolers.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 95–824).

Wolter, K. 1985. *Introduction to Variance Estimation.* New York: Springer-Verlag.

Chapter 4

Nonresponse Error

4.1 Introduction

Nonresponse is an error of nonobservation like coverage error (see chapter 5). However, nonresponse error differs from coverage error in that nonresponse reflects an unsuccessful attempt to obtain the desired information from an eligible unit, whereas coverage error (as noted in the next chapter) reflects the failure to have the sample unit uniquely included in the frame.

A great deal of research and attention has been devoted to nonresponse errors (for reviews see Groves 1989; Groves and Couper 1998; Lessler and Kalsbeek 1992). During the last decade, issues of nonresponse error and nonresponse bias in survey estimates have become an important federal interagency survey research topic. In 1997, two interagency work groups, one addressing household surveys and the other addressing establishment surveys, began discussions that focussed specifically on the problem of nonresponse error (Atrostic and Burt 1999; Kydoniefs and Stanley 1999). A previous subcommittee of the FCSM was also devoted to nonresponse in federal surveys; however, that group chiefly focused on trends in nonresponse rates over time (Gonzalez, Kasprzyk, and Scheuren 1995; Shettle et al. 1994; Johnson et al. 1994; and Osmint et al. 1994). Before that, a panel on incomplete data established by the Committee on National Statistics (CNSTAT) in 1977 provided recommendations on dealing with nonresponse and published three volumes of case studies and research articles (Madow et al. 1983).

There are two main types of nonresponse error: unit nonresponse and item nonresponse. Unit nonresponse is a complete failure to obtain data from a sample unit, whether the sample unit is a household, person within a household, or a business establishment. Despite the best efforts during data collection, some level of unit nonresponse is likely to occur. Weighting techniques are often used to minimize the effect of unit nonresponse error. However, to the extent that the underlying assumptions are not fully met, such as, for example, the sample units are not missing at random, nonresponse error may still affect estimates derived from the data.

Unit nonresponse has received a great deal of attention, due in part to the frequent use of the survey response rate as an overall indicator of the quality of the data. In fact, it is often the only quantitative indicator available or widely used. Nonetheless, the nonresponse rate is only an indicator of the *potential* bias in a survey estimate due to nonresponse errors in the survey. The actual degree of nonresponse bias is a function of not only the nonresponse rate but also how much the respondents and nonrespondents differ on the survey variable of interest. Thus, the effects of nonresponse errors are very rarely directly observed due to the difficulty in obtaining information from and about the nonrespondents. Nonetheless, there are methods and special research studies that can be done to provide some information about the degree nonresponse error may affect an estimate.

Because the amount of nonresponse error is difficult to measure, efforts are often directed to minimize its occurrence. As a result, measures of the processes followed in data collection to minimize nonresponse may also serve as indicators of data quality.

Item nonresponse occurs when a responding unit does not complete an item or items on the survey questionnaire, or the response(s) obtained are unusable. Item nonresponse also affects the quality of the data. Recontact and followup efforts may be used to improve item response rates, particularly on critical data items. Measures of these efforts may also serve as indicators of data quality. Remaining item nonresponse may be dealt with by using imputation methods to compensate for the item nonresponse. These methods do not wholly compensate for the missing data, resulting in some unknown level of error remaining in the data due to nonresponse.

It is important to note that unit and item nonresponse are not completely distinct. If a survey questionnaire has been returned with responses on only some items, but not the critical items, it may be treated as a unit nonresponse. Furthermore, a mail survey questionnaire returned with no items answered may be treated as unit nonresponse, even though the lack of information on the questionnaire makes it impossible to determine whether the sample unit is in- or out-of-scope.

This chapter contains a discussion on reporting nonresponse error, including the indicators of quality pertaining to nonresponse. For unit nonresponse and item nonresponse we discuss separately the calculation of response rates, studies of nonresponse bias, and methods to compensate for nonresponse. We also discuss the processes and procedures used to minimize both unit and item nonresponse. The chapter concludes with a discussion of the reporting of nonresponse in federal surveys.

4.2 Unit Nonresponse

A variety of reasons exist for unit nonresponse, and they may vary depending on the mode of the survey and the survey design. Unit nonresponse in an interviewer-administered personal visit household survey can occur because no one is home (noncontact), refusal to participate, or inability to participate due to language barriers or cognitive or physical incapacity to respond. Sample persons in a telephone survey may refuse to participate directly, or may be using answering machines and caller ID to screen calls, and thus, in some sense refuse to participate without being formally solicited. Sample persons in mail surveys may refuse by failing to return the survey form. In this situation it can be difficult to distinguish noncontacts or nonreceipt of a survey questionnaire from refusals. Similarly, in establishment surveys, the survey questionnaire may never reach the appropriate contact person or the survey form may be sent to the wrong location, resulting in unit nonresponse.

It is important to identify and measure the different reasons and components of nonresponse because different levels of nonresponse bias may be associated with different reasons for nonresponse. For example, noncontacts, those who were never given the opportunity to choose whether or not to participate in the survey, may have very different characteristics than refusals, who were contacted but chose not to participate, and both may differ from survey respondents on some survey variables. Very different trends over time may exist for some of these components, and these trends should be monitored. For example, the increasing use of answering machines, caller ID, and call blocking devices may result in more noncontacts in surveys that use telephone and personal visits, but may appear only as cases with undetermined eligibility in surveys that use only telephone.

To define unit nonresponse for a survey it is first necessary to define the unit of analysis of the survey. Some survey designs are hierarchical and several possible levels or stages of nonresponse are possible before reaching the ultimate sampling unit. For example, an education survey that samples classroom teachers may first sample school districts, then schools, and finally teachers. Nonresponse may occur at the district level or the school or individual teacher levels. Similarly, hierarchical designs occur in health surveys where nursing homes or hospitals are first sampled, and then in the second stage of sampling patients are sampled. Nonresponse may occur at both stages of sampling. Because of the cumulative effects of nonresponse in hierarchical surveys, even rather high levels of response at each level can still result in relatively low response rates at the ultimate sampling unit level.

Longitudinal survey designs commonly have panel nonresponse, which is unit nonresponse to one or more waves of the survey. In long-term longitudinal surveys, panel nonresponse is often permanent and results in attrition of cases from the sample, i.e., very few cases drop out one time and return the next. In this case, the effective level of nonresponse increases over time.

4.2.1 Computing and Reporting Response/Nonresponse Rates[1]

As noted earlier, the response rate or its complement, the nonresponse rate, is commonly used as an indirect measure of the quality of survey data. The computation of response rates should be straightforward: the number of responding units divided by the total number of eligible units. However, there are many different ways of calculating response and nonresponse rates, and efforts to standardize these (Council of American Survey Research Organizations 1982; American Association for Public Opinion Research 2000) have not met with complete success. American Association for Public Opinion Research (2000) does, however, try to provide a general framework for response rates that can be applied across different survey modes. Smith (forthcoming) furthers this discussion by examining the development of definitions and standards concerning nonresponse rate calculations.

One difficulty inherent in response rate calculations is that the eligibility of a nonresponding case may sometimes be difficult to determine. For example, it can be very difficult to discern whether telephone numbers that are never answered are nonworking numbers versus households with no one at home. Similarly, mail surveys not returned by the postmaster are assumed to have reached the correct household or business establishment, but they may not have been received.

4.2.2 Unweighted Response Rates

As stated above, the basic definition of the survey response rate is the number of interviews with reporting units divided by the number of eligible reporting units in sample. The calculation of this rate often varies depending on the treatment of partial interviews and the treatment of sample units of unknown eligibility. The discussion that follows has benefited from the work of a committee of the American Association for Public Opinion Research (2000).

To illustrate the calculation of some common response rates, it is necessary to define the possible outcomes. For the purpose of this discussion, all partial interviews have been reclassified as a completed interview or a noninterview depending on the extent to which the questionnaire was completed. With this simplification, survey cases can be categorized into four groups: interviewed cases, eligible cases that are noninterviews, cases of unknown eligibility, and cases that are not eligible.

[1] This section draws heavily on work done by a previous FCSM subcommittee on nonresponse.

We let:

I = number of interviewed cases
R = number of refused interview cases
NC = number of noncontact sample units known to be eligible
O = number of eligible sample units, noninterviewed for other reasons
U = number of sample units of unknown eligibility, no interview
e = estimated proportion of sample unit cases of unknown eligibility that are eligible

The most useful and simple response rate for surveys in which sample units have an equal probability of selection and the frame includes no ineligible units is:

$$RR1 = \frac{I}{I + R + O + NC} \tag{4.1}$$

This response rate is the one most often used in interviewer-administered federal household surveys. For example, the Current Population Survey (CPS), the National Crime Victimization Survey (NCVS), the Consumer Expenditures (CE) Survey, and the National Health Interview Survey (NHIS) conducted by the U.S. Bureau of the Census all report response rates using this formula.[2]

When the sample frame is imperfect and some cases have uncertain eligibility, as is common in random digit-dialing (RDD) telephone surveys in which a phone is not answered after repeated attempts, the denominator should be an estimate of the total number of eligible units. Thus, one may compute a response rate using equation 4.1, ignoring the cases, *U*, with undetermined eligibility—and overestimate the true response rate—or include those cases in the denominator—and probably underestimate the true response rate. In this case, surveys may report more than one response rate with the true value somewhere in between. Conversely, they may attempt to estimate the true rate by estimating the number of the cases that would be eligible using an estimate of the probability of a case being eligible given that it is undetermined.

If we let *e* be the estimated proportion of cases of unknown eligibility that are eligible, then, *e* multiplied by the number of undetermined cases, *U*, as shown in equation 4.2 is the estimated number of sample unit cases eligible for interview from the sample unit cases of unknown eligibility. The proportion, *e*, may be based on results from the current sample or may be based on other research or surveys or guidelines such as those put out by the Council of American Survey Research Organizations (CASRO).

$$RR2 = \frac{I}{I + R + O + NC + eU} \tag{4.2}$$

For example, the National Household Education Survey (NHES) is an RDD survey sponsored by the National Center for Education Statistics that reports four response rates, each of which uses a different set of assumptions for the allocation of the cases with unknown residential status (table 4.1). The first response rate uses the proportion of households *identified in checks with telephone*

[2] The U.S. Bureau of the Census also calculates a nonresponse rate that includes the *R*, *NC*, and *O* cases into a category called Type A nonresponse. The Type A nonresponse rate is the complement to the response rate.

business offices to allocate the proportion of cases with unknown residential status (U) in the denominator. This rate was 40.5 percent for the study cited in table 4.1. The second response rate (*CASRO response rate*) uses the proportion of households observed in the study with known residential status (e=47.2 percent) to allocate the proportion of cases with unknown residential status (U) in the denominator. These rates are both based on equation 4.2. The difference between the response rates in table 4.1 is due to the estimate used for e. The other two response rates in table 4.1 are a conservative response rate, which includes all cases (U) of unknown eligibility as eligible (that is, e=1) and a liberal response rate, which excludes all cases of unknown eligibility (that is, e=0) (table 4.1).[3]

Table 4.1.—Example of reporting multiple response rates for random digit-dialing (RDD) surveys: 1996 National Household Education Survey

Screener response category	Number	Percent of all numbers	Percent of residential numbers
Total	**161,446**	**100.0**	
Identified as residential	76,258	47.2	100.0
Participating	55,838	34.6	73.2
Not participating	20,420	12.6	26.8
Identified as nonresidential	75,736	46.9	
Unknown residential status	9,452	5.9	
Screener response rates*		**Rate (Percent)**	
Estimated response rate (using business office method)		69.9	
CASRO response rate (using known residential data)		69.1	
Conservative response rate (using all U, e=1)		65.4	
Liberal response rate (using no U, e=0)		73.2	

* All of the response rates use the weighted number of participating households as the numerator (see table 1 from Montaquila and Brick 1997, 17).

SOURCE: Montaquila, J. and Brick, J. M. 1997. *Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey.* Washington, DC: U.S. Department of Education. National Center for Education Statistics (Working Paper No. 97–40).

When the sample design is hierarchical, nonresponse can occur at any stage, e.g., a school survey that selects schools and teachers within schools. In this case, one can estimate the proportion of eligible units that were not selected because of nonresponse at a higher level using equation 4.2.

[3] The NHES publications and technical reports typically report all of these as weighted response rates, which are described below; however they are also calculated as unweighted response rates to monitor data collection.

Alternatively, one can compute response rates at each level using equation 4.1 and then calculate the final response rate by multiplying the rates at each level. For example, the response rates for the NCES' Teacher Survey from the Schools and Staffing Survey (SASS) are calculated and reported in this manner (table 4.2).[4]

Table 4.2.—Hierarchical response rates: Response rates for the Schools and Staffing Surveys, public schools 1993–94

Public school component	Sample size	Unweighted response rate	Weighted[1] response rate	Weighted overall[2] response rate
Local Education Agency (School District) Survey	5,363	93.1	93.9	93.9
Administrator	9,415	96.6	96.6	96.6
School	9,532	92.0	92.3	92.3
Teacher	53,003	88.9	88.2	83.8
Library	4,655	91.1	90.1	90.1
Librarian	4,175	93.5	92.3	92.3
Student	5,577	90.2	91.3	80.3

[1] Weighted using the inverse of the probability of selection.

[2] The weighted overall teacher response rate is the product of the weighted response rate of teachers and the response rate of schools providing a list of teacher (95 percent). The weighted overall student response rate is the product of the weighted student response rate and the response rate of schools providing a list of students (88 percent).

SOURCE: Excerpted from Monaco, D., Salvucci, S., Zhang, F., Hu, M., and Gruber, K. 1998. *An Analysis of Total* N*onresponse in the 1993–94 Schools and Staffing Survey (SASS).* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 98–243).

4.2.3 Weighted Response Rates

It is common for household and establishment surveys to have different probabilities associated with different units in the sample in order to reduce the variance in subpopulation estimates or estimates for the total population. In this case, it is useful for the response rate to reflect the selection weight, w_i . For each observation i, let $I_i = 1$ if the *ith* respondent is an interview, and $I_i = 0$, if the *ith* respondent is not an interview. Similarly, let $R_i = 1$ if the *ith* sample unit is a refusal, and $R_i = 0$ if the *ith* sample unit is not a refusal; $NC_i = 1$ if the *ith* sample unit is a noncontact sample unit known to be eligible, and $NC_i = 0$ if the *ith* sample unit is not a noncontact sample unit known to be eligible; and $O_i = 1$ if the *ith* sample unit is a noninterview for reasons other than a refusal and known to be eligible, and $O_i = 0$ if the *ith* sample unit is a noninterview for reasons other than a refusal and known to be eligible.

[4] The SASS Teacher Survey response rates are also calculated both weighted and unweighted.

A weighted response rate, with w_i = the inverse of the probability of selection for the *ith* sample unit and the sum is over all sample units selected to be in sample, can be given by:

$$\frac{\sum w_i I_i}{\sum w_i (I_i + R_i + NC_i + O_i)} \qquad (4.3)$$

Equation 4.3 can be modified to reflect cases of unknown eligibility. This response rate shows the estimate of the proportion of the population measured if the same survey procedures were used on a census of the population. This rate may be useful for characterizing how completely the population was measured prior to adjustment for nonresponse and can also be useful for showing separate response rates for subpopulations that had different selection probabilities.

In addition, many establishment surveys use weighted response rates because a small number of extremely large firms may dominate an industry (Osmint, McMahon, and Ware-Martin 1994). Because nonresponse by one of these large firms can have severe consequences on how well a variable, such as total sales, is estimated, weighted response rates in establishment surveys are typically reported as an estimate of the proportion of the population total for a characteristic, *y*, associated with the responding sample unit. The characteristic, *y*, for example might be “total assets” or “total revenues” or the “total amount of coal produced.” This response rate is often called a “coverage rate” but, in fact, is a weighted item response rate, where the item of interest is a quantity of primary interest for the survey. If we let y_i be the value of the characteristic *y* for the *ith* sample unit and sum over the entire sample, then the weighted response rate is given as:

$$\frac{\sum w_i y_i I_i}{\sum w_i y_i (I_i + R_i + NC_i + O_i)} \qquad (4.4)$$

Equation 4.4 can be modified if there are cases of unknown eligibility. Data collected for a previous period or from administrative records may be used in the denominator in a weighted response rate to represent the nonrespondents.

4.2.4 Using Weighted versus Unweighted Response Rates

Both unweighted and weighted response rates can be useful for different reasons in any survey. Unweighted response rates provide an indicator of the quality of the data collection and may be calculated at national, regional, and interviewer levels or for certain domains to evaluate performance. The unweighted nonresponse rate (or the number of nonrespondents) indicates the extent of nonresponse followup—a data collection workload measure. Weighted response rates indicate the proportion of the population (or some calculated subpopulation) that responded to the survey, and can be useful for an analyst’s evaluation of the effect of nonresponse on survey estimates (Kasprzyk and Kalton 1997; Madow, Nisselson, and Olkin 1983). For example, the SASS and NHES surveys (both sponsored by NCES) rely on unweighted response rates during data collection to monitor progress. However, they analyze weighted nonresponse rates to determine whether portions of the population are underrepresented in their surveys due to nonresponse (Jabine 1994; Scheuren et al. 1996). For establishment surveys, the weighted response weight reflects the proportion of some characteristic of interest covered by the sample, and reflects the amount the characteristic will need to be estimated by using imputations or weight adjustments.

4.2.5 Other Response Rates

Longitudinal surveys also present some special concerns for the calculation of response rates. Longitudinal surveys are not only subject to nonresponse at the initial interview as a result of refusals, unable to participate, etc., but also are subject to nonresponse for the same reasons at later interviews. If nonrespondents in one wave of a longitudinal survey are eligible sample units in subsequent waves of a longitudinal survey, response rates corresponding to the missing interview pattern can be calculated (Lepkowski 1989). The response rate in longitudinal surveys, however, often focuses on attrition, i.e., those who are not interviewed again after being interviewed one or more times. For surveys that are person-based, it is common to report a retention rate, which indicates the percentage of the persons interviewed who were interviewed at later waves divided by those eligible for interview. For example, the National Longitudinal Survey (NLS) reports retention rates for each year's interviewing. In this case, it may be difficult to determine the eligibility of some sample persons who are not locatable.

Longitudinal surveys that are household- and person-based have further complications because households may leave the survey universe or split up after the first interview, and new persons may also enter established households. All of these changes require that more persons and households be located and tracked over time. Thus, the exact number of eligible households or persons may not be known from nonresponding or unlocatable households. For example, in the Survey of Income and Program Participation (SIPP), Lepkowski, Kalton, and Kasprzyk (1989) estimated the number of nonresponding persons from the number of nonrespondent households by multiplying the average number of sample persons in responding households by the number of nonresponding households.

It may also be useful to calculate other response rates for specific purposes. For example, in a mail survey one may want to report a completion rate—the ratio of completed questionnaires to all questionnaires mailed out. This rate would be the same as equation 4.1 except that one would include the cases with undetermined eligibility (*U*) as well as ineligible or out-of-scope cases in the denominator. In addition, economic surveys that release preliminary estimates before all data are collected may provide response rates for the preliminary and final stages and any others in between.

It may be useful to distinguish between refusals and noncontacts to identify different kinds of problems in field procedures. One may compute a contact rate to see how many households are not being reached, as well as a cooperation rate, which reflects the percentage of contacted cases that actually participate. A high noncontact rate may indicate lack of interviewer effort, or poor timing, while a low cooperation rate may reflect the need for better training in selling the survey or stronger efforts on refusal conversion.

It is also common to calculate response rates at various levels to monitor data collection and performance of interviewers. For example, the U.S. Bureau of the Census typically calculates response rates for each regional office for many of its surveys. In turn, the regional offices calculate response rates for each interviewer or team of interviewers for particular surveys. Most data collection organizations include response rates as a component (sometimes the major component) in their evaluation of interviewers.

Finally, surveys that use different modes of data collection, such as initial mailings followed by telephone interviews with nonrespondents, with final followup by personal visit for difficult or reluctant respondents, may want to calculate the response rates for each mode as well as an overall response rate. This allows for the analysis of data quality associated with individual modes of collection. For example, in the Current Population Survey (CPS), the U.S. Bureau of the Census calculates a recycle rate for cases that are sent to a centralized telephone center for interviewing. Cases that are not contacted by the telephone center or refuse to participate are sent back (or "recycled") to the field interviewer for followup. A response rate for field interviewer followup cases would be informative as well.

4.2.6 Measuring and Reporting Nonresponse Bias

As noted earlier, the degree of nonresponse error or nonresponse bias is a function of not only the nonresponse rate but also how much the respondents and nonrespondents differ on the survey variables of interest. For a sample mean, an estimate of the bias of the sample respondent mean is given by:

$$B(\bar{y}_r) = \bar{y}_r - \bar{y}_t = \left(\frac{n_{nr}}{n}\right)(\bar{y}_r - \bar{y}_{nr}) \qquad (4.5)$$

where:

$\bar{y}_t$ = the mean based on all sample cases
$\bar{y}_r$ = the mean based only on respondent cases
$\bar{y}_{nr}$ =the mean based only on nonrespondent cases
n = the number of cases in the sample
n_{nr} = the number of nonrespondent cases

$\bar{y}_r$ is approximately unbiased if either the proportion of nonrespondents (n_{nr}/n) is small or the nonrespondent mean, $\bar{y}_{nr}$, is close to the respondent mean, $\bar{y}_r$.

Nonresponse may also result in an increase in the mean square errors of survey estimates, a distortion of the univariate and multivariate distributions of survey variables, and therefore result in biased estimates of means, variances, and covariances. Because of the smaller sample size due to nonresponse, variances of estimators are also increased.

The nature of nonresponse is such that values for nonrespondents on all survey measures are not available. A variety of methodologies have been developed and used to estimate the amount of nonresponse bias for at least some survey variables. Some of these methodologies require conducting special studies, while others take advantage of information available in the sampling frame or information gathered during data collection.

The potential for nonresponse bias can be assessed in what are often called identification studies (e.g., see Lessler and Kalsbeek 1992). In this type of study the characteristics of respondents and nonrespondents are compared on a variety of sociodemographic characteristics available from the sampling frame or some external source. If the distribution of the available variables is similar for respondents and nonrespondents, then the concern over nonresponse bias is lessened. In contrast, if the distributions are different the observed differences on the available variables may provide some insight as to the differences that exist on other survey measures. For example, analysis of the school nonresponse rates in the 1988 National Education Longitudinal Study found differences by type of school (public/private) and region of the country, but not by school size or minority enrollment. Furthermore, analysis of 14 key items collected from both responding and nonresponding schools only identified potential bias in three of the key items, all of which were related to low response rates in public schools (Spencer et al. 1990).

One variation of an identification study is to examine the characteristics of respondents to different waves of data collection, most commonly in mail surveys. In a study of respondents to the 1993 survey of Doctorate recipients, Moonesinghe, Mitchell, and Pasquini (1995), examined the weighted responses of respondents who were "early" (responded to the first wave mailing), "interim" (responded to the second or first mailing), and "late" (interim plus CATI followup interviews). They found that those who were academically employed, teaching, and full professors may be under-represented among the "early" respondents, as compared to the "interim" or "late" estimates, thus potentially biasing the "early" estimates.

A variation of this approach can be found in Tucker and Harris-Kojetin (1998) who take advantage of panel survey data from the Current Population Survey. They use information from the previous month in which the household reported to understand the characteristics of nonrespondents and the consequence of nonresponse on labor force estimates. They compare characteristics of households that are respondents in two consecutive months with those that are nonrespondents in one of the two months.

Another variation of an identification study is to match a sample of survey respondents and nonrespondents with a complete data source, such as a census. For example, Groves and Couper (1998) report results from six federal household interviewer-administered surveys that they matched with 1990 decennial census data to learn about the demographic characteristics of those who did not participate in large-scale, face-to-face interviews. Using household and block-level information from the census, they developed distinct multivariate models for households that were contacted compared to those not contacted, and households that cooperated compared to those that were contacted, but did not cooperate. In the case of contact rates, availability and access seem to play a role. Single family homes, households with young children, and the elderly tend to have higher contact rates; while households in multiunit structures, single person households, and households in urban areas tend to have lower contact rates. Cooperation rates tend to be higher in older households, households with young children, and households with young adults; but those who live alone tend not to cooperate.

The 1987–88 National Food Consumption Survey (NFCS) compares demographic characteristics of respondents with estimates from the Current Population Survey to establish indirect inferences regarding nonresponse bias that may result if the unweighted NFCS data were used for analysis (Guenther and Tippett 1993). These analyses are part of a larger body of work described in Guenther and Tippett (1993) in which comparisons on other contemporaneous surveys and past surveys are undertaken to establish the extent of bias due to nonresponse. In addition, because the NFCS consisted of several survey components, intrasurvey comparisons were made on the same items across components to look for potential patterns of nonresponse bias.

Random digit dial (RDD) surveys typically have little or no frame information, making it difficult to compare respondents and nonrespondents. However, the National Household Education Survey (NHES) mapped their RDD telephone exchanges onto zip codes and used census information about households in those zip codes to compare areas with higher and lower response rates and, thus, obtain some indication of possible nonresponse bias (see Montaquila and Brick 1997).

Another methodology that can be used is to subsample nonrespondents to a survey and use extensive followup and conversion procedures to obtain complete cooperation from that subsample. Survey statistics can then be calculated for the subsampled nonrespondents and used to estimate parameters for all of the nonrespondents. The National Education Longitudinal Survey (NELS:88) used these techniques between the base year and the first followup to improve the estimates for students with disabilities and for students with limited English proficiency (Ingels 1996).

4.3 Item Nonresponse

As noted earlier, item nonresponse occurs when a respondent provides some, but not all, of the requested information, or if the reported information is not useable. The line between item and unit nonresponse is sometimes not clear. If a respondent breaks off an interview or sends in a partially completed questionnaire, at what point is it considered item nonresponse versus unit nonresponse? To the extent that survey managers draw the line at different points, there can be differences in the level of unit nonresponse as the level of item nonresponse that is tolerated changes. For example, if a particular survey manager requires 90 percent of all items to be answered for a completed questionnaire, it is possible that a number of partial interviews would be treated as unit nonresponse. But if a different survey manager changed the required level of item response to 80 percent, the number of partial interviews treated as unit nonresponse would decrease and the unit response rate would increase.

4.3.1 Causes of Item Nonresponse

Item nonresponse may occur for a variety of reasons and in a variety of ways (Groves 1989; Madow, Nisselson, and Olkin 1983). A number of items may be missing because the respondent broke off the interview after partially completing it, but enough data were provided so that the questionnaire is not classified as a unit nonresponse. Single or multiple items may be missing because the respondent inadvertently skips an item or block of items on a self-administered questionnaire or refuses to answer the question(s). Sometimes, the respondent may not have the information to answer the question, and this may occur more frequently when the respondent is a proxy for another person. However, Groves (1989) also notes that “don’t know” is a meaningful response for some questions such as those inquiring about opinions of a political candidate or a referendum. Finally, items may be missing because the interviewer skips an item or block of items or fails to record a respondent’s answer.

The greatest concern with the respondent’s refusal to provide the requested data is that there may be systematic differences between those who provide the information and those who do not. Respondents may have good reasons for refusing to answer certain items. Item-missing rates are frequently highest for issues people perceive as private or sensitive, such as income. Items that require greater cognitive effort or confuse the respondent may also have higher nonresponse rates. Because item-missing data occur after the respondent agrees to participate, item nonresponse rates can also be important quality indicators that reflect problems in survey question wording or response options. In other words, item-missing data also reflect measurement errors as well as nonresponse errors (Groves 1989), and many of the issues discussed in chapter 6 are quite relevant to item nonresponse. For example, the quality profile of the SASS (Jabine 1994) describes research relating problems in the design and layout of questions to high item nonresponse rates. Changes were made to some items and the differences in nonresponse rates between rounds 1 and 2 of the survey showed considerable improvement in item nonresponse due to these design changes.

Nonresponse rates for specific items or sections of a questionnaire are an important quality indicator for data items in the same way that the unit nonresponse rates are used as overall indicators of the quality of survey data. Furthermore, agencies may set requirements at certain levels of overall response prior to developing the data for analysis, publication, and public use datasets.

4.3.2 Computing and Reporting Item Response/Nonresponse Rates

Item nonresponse rates are typically computed only for responding units. The item nonresponse rate is usually calculated as the ratio of the number of eligible units responding to an item to the number of responding units eligible to have responded to the item (Madow, Nisselson, and Olkin 1983). It may be useful to calculate item nonresponse rates for specific important items in a survey or for blocks of questions in the questionnaire or an overall level for the entire questionnaire. For example, the quality profile for the American Housing Survey shows the item nonresponse rates for 43 selected items (Chakrabarty and Torres 1996).

It is also useful to report item nonresponse rates separately for different kinds of respondents or by important survey variables. For example, the National Science Foundation's Survey of Graduate Students and Postdoctorates in Science and Engineering reports item-missing rates[5] separately for different subject areas, for example, Computer Sciences, Biological Sciences, as well as for full-time and part-time graduate students and postdoctoral students (National Science Foundation 1999).

As with unit nonresponse rates, item nonresponse rates can be computed at different levels. It can be useful to calculate item-missing rates for interviewers or groups of interviewers to monitor the quality of the data they are collecting. In a household survey, it may be useful to compute item-missing rates separately for the whole household as well as for individuals within the household.

4.4 Compensating for Nonresponse

Given that a survey has encountered some level of nonresponse, the survey manager is faced with the problem of how to compensate for it. Although there are several options including doing nothing and working harder to decrease the number of nonrespondents (see Lessler and Kalsbeek 1992, for several alternatives), typically some method is used to attempt to compensate for nonresponse. There are two general approaches to compensating for nonresponse, adjustment as part of the estimation process (e.g., weighting adjustments) and directly estimating the value each nonrespondent might have reported (imputation) had he/she been a respondent. Kalton and Kasprzyk (1986) note that weighting adjustments are primarily used to compensate for unit nonresponse while imputation procedures are more likely to be used to compensate for missing items. For some periodic surveys, such as those conducted by the Energy Information Administration, imputation procedures are used to estimate all survey items for unit respondents.

[5] The item-missing rates are reported as imputation rates, because missing items are imputed, see section 4.2.3.

4.4.1 Weighting Procedures

A variety of different weighting procedures and models can be used, e.g., population weighting, sample weighting, ratio, and response propensity. See, for example, Kalton and Kasprzyk (1986), Lessler and Kalsbeek (1992), Oh and Scheuren (1983), Brick and Kalton (1996) for summaries of different procedures. The purpose of all these procedures is essentially to increase the weights of the respondent cases to represent the nonrespondents.

Nonresponse weighting adjustments require some information about the nonrespondents and respondents as well as assumptions about the differences between respondents and nonrespondents. Except for special studies (as described in the previous section) little is known about nonrespondent cases other than information that is contained on the sampling frame. For example, in the Current Population Survey, the nonresponse adjustment is based only on clusters of similar sample areas (using Metropolitan Statistical Area (MSA) status and size) that usually are contained within a state. Each MSA cluster is further split into "central city" and "balance of MSA," while non-MSA clusters are split by "urban" and "rural" residence.

In the case of panel studies, data from the initial interview can be used to develop nonresponse adjustments to the weights for future waves of interviews. Reporting on the information used to adjust for nonresponse can be valuable for analysts who wish to study alternative methods of adjustment or evaluate the utility of the adjustment procedures. Lepkowski, Kalton, and Kasprzyk (1989) used longitudinal data from the SIPP to examine the characteristics of panel nonrespondents and evaluate alternative nonresponse weighting adjustments. They found that the majority of nonrespondents were interviewed at least once, and thus a great deal was known about them from the interview(s) they gave. After finding significant differences between complete and partial respondents on a variety of demographic characteristics, they used alternative modeling techniques to develop new weight adjustments. They compared and evaluated the effects of the different weights with the initial panel weights by examining how highly correlated they are with each other and by comparing survey estimates produced with each set of weights. As a final step, they compared weighting procedures by examining the distribution of weights that they produce. The more variable the weights are, the greater the loss of precision in the survey estimates. Overall, their analyses did not show the alternative weighting schemes to be more effective in compensating for panel nonresponse than the method being used.

Harris-Kojetin and Robison (1998) used panel data from the Current Population Survey to examine the effects of unit nonresponse. Since most CPS nonrespondents are actually respondents for at least one month when in sample, data obtained from other months can be used both to compare respondents and nonrespondents as well as to evaluate the unit nonresponse adjustment procdures.

Because weighting for nonresponse can affect the quality of the estimates from a survey, it is important for agencies to research the effects of nonresponse adjustments on statistical estimates. To evaluate nonresponse adjustment procedures typically requires some external source of data similar to that used in nonresponse bias studies. For example, a National Center for Health Statistics study revealed that a weighting procedure used during the 1988 cycle of the National Survey of Family Growth may have reduced nonresponse bias. The researchers compared estimates of the number of United States births resulting from the 1988 weighting procedure with administrative records and found the procedure produced more accurate estimates than the 1982 weighting procedure (see Mosher, Judkins, and Goksel 1989).

As noted in section 4.1, it may be important to distinguish between refusals, noncontacts, and other noninterviews because the underlying causes of each type of nonresponse may be quite different. Nonresponse adjustment procedures have typically not taken into account the different kinds of nonresponse, but, in an extensive study, Groves and Couper (1998) match sample survey data to the 1990 U.S. Decennial Census to create separate models of cooperation and contact from six federal surveys. They suggest how these models could be used to improve nonresponse adjustment procedures. Given the differences they found between models for noncontact and refusals, they suggest that the source of nonresponse and the related correlates should be taken into account when response propensity models are used to develop nonresponse weight adjustments.

Nonresponse adjustment models can only incorporate information obtained on both respondents and nonrespondents. This has resulted in recommendations that further efforts be devoted to obtaining additional information on nonrespondents as part of the survey design to enrich and improve the nonresponse adjustment models (Groves and Couper 1995; Madow, Nisselson, and Olkin et al. 1983).

4.4.2 Imputation Procedures

Typically, a great deal more data are available to examine the impact of bias due to item nonresponse than unit nonresponse because respondents have answered a sufficient number of items in the survey to be considered at least a partial interview. As a result, comparisons can be made between respondents and nonrespondents on the items that were completed to attempt to infer whether there are any systematic differences. Frequently, the data from the items that were answered are used to model or impute values for those that are missing.

Survey managers must decide how to deal with item-missing data. One could do nothing and conduct analyses of data on only complete cases. This results in the loss of some usable data and a reduced sample size. However, using only complete cases is not an option if the primary focus of the survey is to provide aggregates such as total sales or total revenues. Alternatively, one could use a similar strategy to unit nonresponse and weight the respondents to adjust for the item nonresponse. However, this could require a different weight for each survey item with missing data. Thus, it is most common to impute for item-missing data rather than weight (Kalton and Kasprzyk 1982).

There are a wide variety of procedures used to impute for item missing data including hot-deck methods, regression methods, and mean imputation (For more information, see Brick and Kalton 1996; Kalton and Kasprzyk 1982 and 1986; Lessler and Kalsbeek 1992; Little and Rubin 1987.). Although imputation methods yield data values for missing items, they come at a cost. Imputation may distort measures of the relationships between variables, and they also distort standard error estimation. There are advantages and disadvantages to each of the different imputation methods. Multiple imputation (Rubin 1987 and 1996) or the assigning of more than one imputed value to a missing data item has been discussed widely during the last 20 years. A major advantage of multiple imputation (Herzog and Rubin 1983; Little and Rubin 1987) methods is that they allow estimates of standard errors to be computed by using the same imputation procedure several times. A disadvantage is that analysts are not accustomed to using multiple imputations in their analyses. The National Center for Health Statistics has developed a data set (available Summer 2001) for the National Health and Nutrition Examination Survey (NHANES) that includes five imputations with detailed instructions for using the data.

Because each imputation method relies on some model (implicit if not explicit) for generating values for the missing items, it is useful to report the method of imputation and variables used in the imputation model, or to identify donor records, or define imputation classes. For example, the American Housing Survey uses prior year data for some items and uses a sequential "hot deck" procedure for other items. The variables used to define the imputation cells vary, but to impute income they use age, race and sex of the person, relationship to reference person, and value of property/monthly rent (Chakrabarty and Torres 1996).

In addition, indicator or shadow variables for each variable that was imputed can be valuable for analysts who wish to use some other form of imputation or analyze only complete records (see section 6.2.4, for example). Indicators of how often a record was used as a "donor" in hot deck imputation can also serve as a quality indicator of the imputation process.

Because imputation for item nonresponse can affect the quality of the estimates from a survey, it is important for agencies to research the effects of those methods. There are several basic methodologies for evaluating the effect of imputation on the data. One way of determining the effect of imputation is to compare estimates that include the imputed values with estimates based only on reported data. Another way of determining the effect of imputation is to compare the estimates to external sources. For example, the CPS hot deck imputation procedure for income was evaluated by using data from the March 1981 CPS income supplement matched to 1980 Internal Revenue Service (IRS) tax records (David et al. 1986).

Agencies can provide in their documentation and in their public-use datasets a great deal of information useful to analysts in evaluating the effects of imputation. For example, the methodology report for the 1992–93 National Postsecondary Student Aid Study (Loft et al. 1995) includes a technical appendix that describes the imputation procedures used for variables that required imputation. Also included are comparisons of the pre- and post-imputation values for the variables.

4.5 Methods and Procedures to Minimize Unit and Item Nonresponse

Federal statistical agencies have used a variety of methods to minimize both unit and item nonresponse, with the goal of improving data quality. The extent of activity to minimize nonresponse can also be an indicator of data quality, and descriptions of these efforts can provide the data user with useful information. Most recurring publications mention that there is followup for nonresponse, but typically do not contain a detailed discussion of how unit or item nonresponse is minimized. Technical and methodological reports provide more detail. An example is the U.S. Energy Information Administration's *Quality Profile of the Residential Energy Consumption Survey* (U.S. Energy Information Administration 1996). The quality profile describes the use of monetary incentives to improve completion rates of automobile use diaries.

There is extensive literature on minimizing nonresponse. See, for example, the report concerning establishment surveys from the Federal Committee on Statistical Methodology (1988) and more recently, Paxson, Dillman, and Tarnai (1995). A number of examples for household surveys have also been presented (Dillman 1978; Groves 1989; Lyberg and Dean 1992). The methods can be grouped into three categories: front-end techniques to promote cooperation, special arrangements, and followup techniques. Recently, some agencies have considered shifting their focus from followup to front-end techniques.

Front-end Techniques. There is no empirical evidence that a single technique produces high response rates. Further, different surveys require different techniques. Studies have shown that positive effects are obtained when using a combination of techniques for some surveys.

Techniques that apply to surveys conducted by interviewers and that are self-administered (both household and establishment surveys) include advance notification in the form of a letter or phone call, use of priority mail, personalization of correspondence, and showing respondents how the data they are providing are being used. Techniques specific to surveys conducted by interviewers include performance guidelines for interviewers and the monitoring and observation of interviewer performance. Techniques specific to self-administered surveys are the use of respondent-friendly questionnaires, that is, questionnaires that are designed to appear easier and less time-consuming to complete, and the use of stamped-return envelopes. Incentives, both financial and nonfinancial, may also help increase response rates. Many in the survey research community believe incentives are an important tool to improve response rates, particularly with a difficult-to-interview population. Kulka (1995) provides a review of the research literature on incentives and a description of current practices.

Special Reporting Arrangements. An example of a special reporting arrangement for small companies is the use of sample rotation so that companies do not have to report each time the survey is conducted, or the use of a shorter version of the form by special arrangement.

Followup Techniques. The Federal Committee on Statistical Methodology (1988) reports that over three-fourths of agency surveys included in their study use intensive followup of critical units and nearly all use some type of nonresponse followup procedure. Followup techniques include reminder cards, periodic telephone calls, and automated fax reminders. Followup techniques help increase the overall response rate; however, if left unmanaged, production schedules can be delayed and the timeliness of the data products affected. Dillman (1991) recommends an integrated approach for front-end and followup techniques so that they are not duplicative.

4.6 Quality Indicators for Nonresponse

Nonresponse errors are likely to be present to some degree in any survey, and they may affect the quality of the data and estimates produced from the survey. The exact amount of nonresponse error on a specific estimate is almost never known. However, there are a number of indicators of the quality of those data that can be extremely useful to analysts, customers, and consumers in evaluating the quality of the data and their usefulness.

Frequently the response rate (or its complement, the unit nonresponse rate) is used as an overall indicator of the quality of the data in the survey. There are also a variety of other more specific indicators that provide insights into the quality of the data and the data collection process. Specifically, rates of refusals and noncontacts can be useful when computed for the entire data collection as well as at the level of field administrative units or interviewers. Also, address not locatable, postmaster returns, and undetermined eligibility (e.g., answering machine—eligibility unknown) can be important in monitoring and evaluating mail and telephone surveys.

Some similar quality indicators can be calculated for item nonresponse rates. These provide a more micro-level data quality indicator for a specific estimate or set of estimates, but also may be informative about the data collection process when computed at different levels.

Additional indicators can be useful in evaluating the potential for nonresponse error in a survey by focusing even more on the data collection process itself. Documenting the procedures that were followed to minimize nonresponse during the data collection may provide a user with helpful information to evaluate the likelihood of nonresponse errors. For example, quality indicators may include the number of contacts attempted for noncontact cases, the attempts and techniques used to convert refusals, and the number of and kinds of reminders and replacement questionnaires sent for mail surveys. One may have more confidence in the results from a survey that had a lower than desired response rate if systematic and extensive procedures were followed to minimize nonresponse.

Indicators of the amount of nonresponse bias for some estimates may be found through special studies of nonrespondents or by using external data. Furthermore, because steps are often taken in the estimation process to compensate for nonresponse by weighting and imputation, effects of these procedures on the quality of the data can be shown by using external data or special studies of nonrespondents. For example, the validity of the weighting or imputation model used can be evaluated by making direct comparisons of the estimates, generated with and without the procedure, to a reliable external data source.

4.7 Reporting Nonresponse in Federal Surveys

Nonresponse is the most visible and well-known source of nonsampling error. Nonresponse rates are frequently reported and are often viewed as the first area requiring study to assess the potential for bias in survey estimates. As discussed above, response rates are a common output of the data collection process and are calculated differently for different purposes. These rates tend to be treated as a proxy for survey data quality—more so than almost any other indicator that may be proposed. Response rates influence the perception of the overall quality of the survey—low response rates suggesting a poor quality survey—and they are usually available.

In the subcommittee's study of the reporting of error sources in analytic publications, Atkinson, Schwanz, and Sieber (1999) found, somewhat surprisingly, that nonresponse was not always identified as a potential error source and response rates were frequently not reported. Of the publications reviewed, 71 percent mentioned unit nonresponse as a possible source of nonsampling error and only 59 percent indicated an overall response rate. Forty-nine percent of the publications that were reviewed mentioned item nonresponse as a source of error, and only 22 percent reported item nonresponse rates. Twenty percent of the publications presented response rates for subgroups.

These findings are very surprising and somewhat disappointing to those who expect a high reporting standard of the federal statistical agencies. Since response rates are perhaps the most easily obtained measure of survey performance and are usually available routinely—without additional computational effort—precisely defined response rates should be reported routinely in analytic survey reports. As with other sources of error, certain kinds of information about nonresponse is not likely to occur in the technical notes section of analytic reports. Studies to estimate the bias due to nonresponse—either by comparing respondents and nonrespondents characteristics on variables available on the sampling frame or by studying a subsample of nonrespondents who have had intensive followups are likely to be reported in special methodological or technical publications. Synthesis reports that bring together a substantial amount of information about a survey are available as user's guides, survey technical documentation, and quality profiles. In each of these types of reports, substantially more material about nonresponse error can be made available to the user.

The discussion of nonresponse error goes beyond quantifying its magnitude; it also includes the precision of the definition, identification of the nonresponse components, and the nature and effects of procedures to compensate for nonresponse. The nature of reporting may depend on the magnitude of the nonresponse problem. For example, using a variable with a high nonresponse rate in a key analysis should warrant some discussion in the report, whereas, if the variable had a low nonresponse rate, no discussion may be necessary. In the course of reviewing errors due to nonresponse, the subcommittee identified a number of aspects of this source of error that need to be addressed when reporting results in an agency analytic report:

- Unit and item nonresponse should be identified as important sources of error.

- Overall unit response rates (weighted using base weights and unweighted) should be provided as well as definitions of the response rates given; this includes providing definitions of the numerator and denominator in the response rate calculation.

- In multistage designs, weighted (using base weights) and unweighted response rates at each interview level should be given, and an overall response rate computed. Assumptions necessary for the response rate calculation should be stated.

- Longitudinal surveys should report separately the response rate for the first wave of the survey, each followup wave of the survey, as well as the cumulative response rate. Reporting other response rates is encouraged; for example, the response rate of sample units responding in all waves of a longitudinal survey may be informative to the longitudinal data analyst.

- Subgroup response rates should be provided if specific subgroups are important to the analysis found in the report.

- Item response rates should be summarized and items with low response rates identified.

- Unit and item nonresponse adjustment procedures—whether they are weighting procedures or imputation methods—should be identified.

- When unit or item nonresponse rates are lower than the agency seems "reasonable," special studies to assess the bias due to nonresponse should be conducted, the results summarized, and the detailed report referenced.

- If available, studies designed to measure potential nonresponse bias should be referenced.

Technical reports and user's manuals should address and document the myriad important details related to nonresponse as a source of error in surveys, including:

- Procedures used to compensate for missing data, both unit and item nonresponse, should be described and the key variables used in the procedures identified. The effects of these procedures (if known) on the estimates should be discussed.

- Evaluations of the missing data procedures, both unit and item, should be conducted, the results summarized, and a detailed report referenced.

- Special studies that aim to understand or measure the bias due to nonresponse should be conducted, summarized, and the detailed report referenced.

- Steps taken to maximize the response rate and the extent of nonresponse followup should be described.

- Subgroup response rates for key subpopulations should be calculated and made available.

- Reasons for nonresponse (refusals, noncontacts, etc.) should be monitored and reported.

References

American Association for Public Opinion Research. 2000. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Ann Arbor, MI: AAPOR.

Atkinson, D., Schwanz, D., and Sieber, W.K. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Atrostic, B.K. and Burt, G. 1999. "Household Nonresponse: What We have Learned and a Framework for the Future." *Seminar on Interagency Coordination and Cooperation*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 153–180.

Brick, J.M. and Kalton, G. 1996. "Handling Missing Data in Survey Research." *Statistical Methodology in Medical Research*. 5: 215–238.

Council of American Survey Research Organizations. 1982. *On the Definition of Response Rates*. Port Jefferson, NY.

Chakrabarty, R.P. and Torres, G. 1996. *American Housing Survey: A Quality Profile*. Washington, DC: U.S. Department of Housing and Urban Development and U.S. Department of Commerce.

David, M., Little, R.J.A., Samuhel, M.E., and Triest, R.K. 1986. "Alternative Methods for CPS Income Imputation." *Journal of the American Statistical Association*. 81: 29–41.

Dillman, D. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley & Sons.

Dillman, D. 1991. "The Design and Administration of Mail Surveys." *Annual Review of Sociology*. 17: 225–249.

Federal Committee on Statistical Methodology. 1988. *Quality in Establishment Surveys*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 15).

Gonzalez, M., Kasprzyk, D., and Scheuren, F. 1995. "Exploring Nonresponse in U.S. Federal Surveys." *Seminar on New Directions in Statistical Methodology*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 23). 603–624.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Groves, R.M. and Couper, M.P. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.

Groves, R.M. and Couper, M.P. 1995. "Theoretical Motivation for Post-survey Nonresponse Adjustment in Household Surveys." *Journal of Official Statistics*. 11: 93–106.

Guenther, P.M. and Tippett, K.S. (eds.). 1993. *Evaluation of Nonresponse in the Nationwide Food Consumption Survey, 1987–88*. Washington, DC: U.S. Department of Agriculture, Human Nutrition Information Services. Nationwide Food Consumption Survey 1987–88 (NFCS Report No. 87-M-2).

Harris-Kojetin, B.A. and Robison, E. 1998. "Evaluating Nonresponse Adjustment in the Current Population Survey (CPS) using Longitudinal Data." *Proceedings of Statistics Canada Symposium 98: Longitudinal Analysis for Complex Surveys.* Ottawa: Statistics Canada.

Herzog, T. and Rubin, D. 1983. "Using Multiple Imputations to Handle Nonresponse in Sample Surveys." In Madow et al. (Eds.) *Incomplete Data in Sample Surveys*. 2: 209–245. New York: Academic Press.

Ingels, S.J. 1996. *Sample Exclusion in NELS:88, Characteristics of Base Year Ineligible Students: Changes in Eligibility Status after Four Years*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 96–723).

Jabine, T. 1994. *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 94–340).

Johnson, A.E., Botman, S.L., and Basiotis, P. 1994. "Nonresponse in Federal Demographic surveys: 1981–1991." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 983–988.

Kalton, G. and Kasprzyk, D. 1982. "Imputing for Missing Survey Responses." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 146–151.

Kalton, G. and Kasprzyk, D. 1986. "The Treatment of Missing Survey Data." *Survey Methodology*. 12: 1–16.

Kasprzyk, D. and Kalton, G. 1997. "Measuring and Reporting the Quality of Survey Data." *Proceedings of Statistics Canada Symposium 97: New Directions in Surveys and Censuses*. Ottawa: Statistics Canada. 179–184.

Kulka, R. 1995. "The Use of Incentives to Survey 'Hard-to-Reach' Respondents: A Brief Review of Empirical Research and Current Research Practices." *Seminar on New Directions in Statistical Methodology*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 23). 256–299.

Kydoniefs, L. and Stanley, J. 1999. "Establishment Nonresponse: Revisiting the Issues and Looking into the Future." *Seminar on Interagency Coordination and Cooperation*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 181–227.

Lepkowski, J. 1989. "Treatment of Wave Nonresponse in Panel Surveys." In D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh (eds.), *Panel Surveys*. New York: John Wiley & Sons. 348–374.

Lepkowski, J., Kalton, G., and Kasprzyk, D. 1989. "Weighting adjustments for partial nonresponse in the 1984 SIPP panel." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 296–301.

Lessler, J. and Kalsbeek, W. 1992. *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Little, R.A. and Rubin, D. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Loft, J., Riccobono, J., Whitmore, R. Fitzgerald, R., and Berkner, L. 1995. *Methodology Report for the National Postsecondary Student Aid Study*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 95–211).

Lyberg, L. and Dean, P. 1992. "Methods for Reducing Nonresponse Rates: A Review." Paper prepared for presentation at the 1992 meeting of the American Association for Public Opinion Research.

Madow, W., Nisselson, H., and Olkin, I. 1983. *Incomplete Data in Sample Surveys*. New York: Academic Press.

Montaquila, J. and Brick, J.M. 1997. *Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (Working Paper No. 97–40).

Moonesinghe, R., Mitchell, S., and Pasquini, D. 1995. "An Identification Study of Nonrespondents to the 1993 Survey of Doctorate Recipients." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 453–458.

Mosher, W., Judkins, D., and Goksel, H. 1989. "Response Rates and Non-response Adjustment in a National Survey." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 273–278.

National Science Foundation. 1999. *Graduate Students and Postdoctorates in Science and Engineering: Fall 1997 (Detailed Statistical Tables)*. Project Officer: J. Burrelli. Arlington VA (NSF 99–325).

Oh and Scheuren, F. 1983. "Weighting Adjustment for Unit Nonresponse." In Madow et al. (eds.) *Incomplete Data in Sample Surveys*. 2: 143–184. New York: Academic Press.

Osmint, J.B., McMahon, P.B., and Ware-Martin, A. 1994. "Response in Federally Sponsored Establishment Surveys." *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association. 977–982.

Paxson, M.C., Dillman, D., and Tarnai, J. 1995. "Improving Response to Business Mail Surveys." In Cox et al. (eds.) *Business Survey Methods*. New York: John Wiley & Sons. 303–316.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D.B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association*. 81: 366–374.

Scheuren, F., Monaco, D., Zhang, F., Ikosi, G., and Chang, M. 1996. *An Exploratory Analysis of Response Rates in the 1990–91 Schools and Staffing Survey (SASS)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 96–338).

Shettle, C.F., Guenther, P.M., Kasprzyk, D., and Gonzalez, M.E. 1994. "Investigating Nonresponse in Federal Surveys." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 972–976.

Smith, T. Forthcoming. "Developing Nonresponse Standards." In R. Groves, (ed.), *Survey Nonresponse*. New York: John Wiley & Sons.

Spencer, B., Frankel, M., Ingels, S., Rasinski, K, and Tourangeau, R. 1990. *National Education Longitudinal Study of 1988: Base Year Sample Design Report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 90–463).

Tucker, C. and Harris-Kojetin, B.A. August 1998. "The Impact of Nonresponse on the Unemployment Rate in the Current Population Survey (CPS). In A. Koch and R. Porst (eds.), *Nonresponse in Survey Research: Proceedings of the Eighth International Workshop on Household Survey Nonresponse.* Mannheim, Germany: ZUMA. 45–54.

U.S. Energy Information Administration. 1996. *Residential Energy Consumption Survey Quality Profile*. Washington, DC: U.S. Department of Energy.

Chapter 5

Coverage Error

5.1 Introduction

When dealing with a census, a survey, or an administrative record system, several questions come to mind. Are any members of the population of interest systematically omitted or underrepresented? Are units omitted (e.g., people, houses, businesses, farms) about which the survey sponsor would like to gather information, thus creating an error of undercoverage? Are elements included in the population of interest that do not belong there? Are the statistics a result of a data set that includes out-of-scope units or units included twice, creating an error of overcoverage? This chapter discusses coverage errors, how they arise, and the methods used to address them.

To discuss coverage error in a meaningful way, one must first define and discuss the concepts of target populations and frame populations. The *target population* is the set of elements about which information is wanted and parameter estimates required (Sarndal, Swensson, and Wretman 1992). The survey designer working with analysts defines the target population while the survey is still in the planning stages. The user might want to know about the entire population of the United States (perhaps more narrowly defined as those who spent 90 of the last 120 days in the United States). Geographic considerations can often be the basis for defining the target population. Other criteria, however, based on subject matter or time, are also used; for example, the target population for household surveys is usually defined as all persons living in the United States in a given period of time and not living in institutions. In an economic application, Cox, Elliehausen, and Wolken (1989) illustrate the difficulties of defining the target population. They discuss the target population of their study in some detail by first providing the general description as “all nonfinancial and nonfarm small business enterprises in the United States in operation as of December 1987,” and then defining the individual concepts included in the population definition.

For practical reasons, the target population will usually differ from the frame population. The *frame population* is the set of all elements that are either listed directly as units in the frame or can be identified through a more complex frame concept, such as a frame for selection in several stages. The frame population is the population of units from which a sample can be selected. Coverage error occurs, for example, when the target population is the noninstitutionalized resident population of the United States, but the frame population is all individuals in the country who have telephones. Coverage error is the difference between the target population and the frame population—that is, individuals living in households that do not have a telephone.

Occasionally, target and frame populations can be defined so that coverage error does not occur. A survey designer may want to obtain data about the entire population of the United States, but may choose to exclude the population of individuals living in nursing homes and hospitals. The Current Population Survey (CPS), for example, explicitly excludes individuals who are institutionalized. Thus, the CPS’s deliberate omission of individuals who are institutionalized is not a coverage error, since CPS excludes such individuals from both the target and frame population. These definitions should be reported and discussed in the survey documentation. In this survey, the exclusion of the institutionalized population is probably a negligible issue for researchers wishing to estimate unemployment rates (which are based on individuals in the labor force). On the other hand, if a household survey were used to estimate the incidence of disabilities among all adults, the exclusion of institutionalized individuals in the frame population would lead to an underestimate of the statistic of interest. Note that deliberate and explicit exclusions to the target and frame population are not considered to be coverage error; however, the identification and reporting of these exclusions is very important to understanding the estimates.

The sampling frame identifies the units from which a sample can be selected, either explicitly or implicitly, and procedures that account for all units of the survey population. A more formal definition of the *sampling frame* is found in Wright and Tsao (1983).

> "The materials or devices which delimit, identify, and allow access to the elements of the target population. In a sample survey, the units of the frame are the units to which the sampling scheme is applied. The frame also includes any auxiliary information (measures of size, demographic information) used for (1) special sampling techniques, such as, stratification and probability proportional to size selections; or for (2) special estimation techniques, such as ratio or regression estimation."

Based on the discussion above, we say that *coverage error* refers to the difference between the target population and the frame population. Coverage error arises from omissions, erroneous inclusions, and duplicates in the sampling frame. *Omissions* reflect the fact that some units in the target population have been omitted from the frame. Omission from the frame means that these units have no chance of being included in the survey. *Erroneous inclusions* reflect the fact that some units not belonging to the target population have been included in the frame. *Duplicates* are defined as target population units that appear in the sampling frame more than once. Omissions give rise to *undercoverage* and erroneous inclusions give rise to *overcoverage*. Kish (1965) provides the traditional definition of coverage error that focuses on one aspect of coverage error, that is, *noncoverage*.

> [N]oncoverage denotes failure to include some units, or entire sections, of the defined survey population in the actual operational sampling frame. Because of the actual (though unplanned and usually unknown) zero probability of selection for these units, they are in effect excluded from the survey results. We do not refer here to any deliberate and explicit exclusion of sections of a large population from the survey population.

Coverage error generally refers to net coverage, the sum of the noncoverage and overcoverage errors. Coverage issues are of concern because units omitted or erroneously included may be different or distinctive in some respect from those included in the survey, thus resulting in biased statistics.

A large literature exists on coverage error and its occurrence. Coverage errors occur because of errors in the sampling frame. This occurs when units are missing or represented more than once, and the frame is not up-to-date; that is, new units have not been added and old units, no longer applicable, have not been deleted. Lessler and Kalsbeek (1992) discuss frame errors and the quantification of frame errors. The Federal Committee on Statistical Methodology (1990) provides a thorough discussion of coverage error and its occurrences prior to and after sample selection; a number of case studies are presented to clarify the issues.

Coverage error differs from another type of nonobservation—namely, nonresponse discussed in chapter 4—and this distinction is not always obvious. Coverage error is totally nonobservable; that is, it leaves no trace of its existence. This is equally true of both undercoverage and overcoverage. The extent of noncoverage can only be estimated against a check outside the data collection procedure itself. This is why this error is difficult to estimate. Nonresponse, on the other hand, results from failing to obtain information on units selected and designated for the sample, due to refusals, failure to locate, etc. The extent of nonresponse can be measured by comparing the selected sample with the achieved sample. Frame variables may also provide information on the characteristics of the nonresponding units. An example may help clarify the difference. Assume that according to survey definitions a structure should be listed as a housing unit, but an interviewer, when developing the sampling frame, is unable to decide whether it is a housing unit or not. If the structure is, in fact, a housing unit, then failure to list it is coverage error—undercoverage. If, however, it is not a housing unit, then listing it is a coverage error—overcoverage. Overcoverage brings with it a chance that the listing will result in nonresponse.

To analyze coverage errors, it is necessary to assume that adjustments have already been made for nonresponse. In the case given above, we assume that the imputation mechanism creates a set of person records for the nonresponse households and these households will have the same coverage as if they had actually responded.

5.2 Measuring Coverage Error

Since coverage errors do not leave any apparent indication of their existence, they can be measured only by reference to an outside source. Coverage error can be measured using indirect and direct techniques. Indirect techniques include comparisons based on the existing sample with comparative data from earlier surveys or from external sources. For example, birth rates of units can be compared from one period to another to indirectly measure coverage error. Similarly, out-of-scope rates, unclassified rates, misclassified rates, and duplication rates provide useful but indirect information on coverage problems (Federal Committee on Statistical Methodology 1988). This section, however, focuses on two direct methods of measuring coverage errors: aggregate comparisons to other sources and case-by-case matching.[1]

5.2.1 Comparisons to Independent Sources

One can sometimes find or construct a better aggregate estimate of the study population than is available from the survey. For example, it is often possible to compare the age, race, and sex distribution of the study population to that of the decennial census, demographic projections made from the census, or estimates based on analytic techniques. Such comparisons must be made taking into consideration errors in both the survey being evaluated and the estimates from

[1] Occasionally, the coverage errors will present indirect internal evidence. For example, in demographic surveys, the age and sex distribution of households by size might suggest certain groups had been omitted. The evidence might equally be the result of classification error. In any case, these specialized demographic techniques will not be discussed here. The interested reader is directed to Shryock, Siegel, and Associates (1975).

the independent source. Differences too large to be attributable to sampling errors may be found, and other sources of error, such as coverage error, must be considered. Of course, the census itself does not have complete coverage, and this should be taken into consideration in the evaluation.

When a better aggregate estimate of the study population is available, a common method of comparing the two estimates is the coverage ratio. The coverage ratio is calculated as the estimate from the survey divided by the "better" aggregate estimate (i.e., an independent population control total) where the survey estimate is first adjusted for nonresponse. An example of the use of a coverage ratio is provided in the accompanying table (table 5.1) which is typical of the information provided in appendices to U.S. Bureau of the Census reports (called "Source and Accuracy of Estimates").

Table 5.1.—Current Population Survey coverage ratios

	Non-Black		Black		All persons		
Age	Male	Female	Male	Female	Male	Female	Total
0–14	0.929	0.964	0.850	0.838	0.916	0.943	0.929
15	0.933	0.895	0.763	0.824	0.905	0.883	0.895
16–19	0.881	0.891	0.711	0.802	0.855	0.877	0.866
20–29	0.847	0.897	0.660	0.811	0.823	0.884	0.854
30–39	0.904	0.931	0.680	0.845	0.877	0.920	0.899
40–49	0.928	0.966	0.816	0.911	0.917	0.959	0.938
50–59	0.953	0.974	0.896	0.927	0.948	0.969	0.959
60–64	0.961	0.941	0.954	0.953	0.960	0.942	0.950
65–69	0.919	0.972	0.982	0.984	0.924	0.973	0.951
70 or more	0.993	1.004	0.996	0.979	0.993	1.002	0.998
15 or more	0.914	0.945	0.767	0.874	0.898	0.927	0.918
0 or more	0.918	0.949	0.793	0.864	0.902	0.931	0.921

SOURCE: U.S. Bureau of the Census. 1992. *Marriage, Divorce, and Remarriage in the 1990's.* Current Population Reports. Washington, DC: U.S. Department of Commerce (P-23–120).

A similar example is reported by Meier and Moore (1999) in which coverage in the National Health Interview Survey is measured by comparing survey estimates after nonresponse adjustment to independently estimated population totals.

The ratio as a measure of coverage has several drawbacks. First, a superior estimate that can serve as a population control must be available or at least constructible; this is not always possible. Second, these comparisons frequently yield only information about net differences in the counts of various groups. For example, one might find that the survey estimated fewer housing units than an outside source indicates are actually present. Aggregate analysis does not give information about the sizes of the gross errors. Errors may "net out" for the categories where independent estimates are available, but still create important bias for other variables of interest. Using an example from business surveys, the estimated number of businesses may be approximately correct; however, if there is overcoverage of large businesses and undercoverage of small businesses, there may be considerable bias in the estimate of average sales. Conversely, if one is interested in estimating total sales within an area, undercoverage of small businesses results in much less bias than undercoverage of large businesses.

The aggregate method, comparing estimates with other data sources, is most informative when it is possible to identify the same subgroups that are undercovered and compute estimates of the relative undercoverage rates. For example, the Natural Science Foundation's Scientists and Engineers Statistical Data System (SESTAT) does not cover foreign-educated scientists and engineers who came to the United States following the last decennial census. However, once a decade, a followup survey of scientists and engineers who came to the United States since the last decennial census provides an estimate of the cumulative undercoverage for this group. One should recognize that differences between a survey and population controls or other independent sources, such as another survey, may be due to factors other than coverage.

5.2.2 Case-by-Case Matching and Dual System Estimation

A second approach to measuring coverage errors is based on case-by-case matching. If an alternative list of units exists or can be constructed, units in the population can be classified as either present or not present in the census/survey/record system as well as on the alternative list.

Consequently, all units can be cross-classified as:

	Alternative frame		
Frame	In	Out	Total
In	$N[11]$	$N[12]$	$N[1*]$
Out	$N[21]$	$N[22]$	$N[2*]$
Total	$N[*1]$	$N[*2]$	$N[**]$

Where the asterisk indicates summation over that column or row. Thus,

$$N[**] = N[11] + N[12] + N[21] + N[22]$$

or, the total population equals the population on both lists, the population on only our list, the population only on the alternative list, and the population not on either list.

The population not on either list is, of course, not observable. However, one can estimate it if the two lists are, approximately, independent. Essentially, this means that the probability of being included on one list does not depend on the probability of being included on the other. See Wolter (1986) for a more precise mathematical description.

Under conditions of independence, we can estimate

$$N[1*] \,/\, N[**] = N[11] \,/\, N[*1]$$

Which is the coverage ratio of our frame, using the alternative list as a control. We can also estimate the total population simply by rewriting this equation

$N[**] = N[1*]\, N[*1] / N[11]$

which is algebraically the same as

$N[**] = N[11] + N[12] + N[21] + N[12]\, N[21] / N[11]$

This estimator has a long history; it has been used in studies of wildlife where it is known as the Peterson estimator (Peterson 1896). In human populations it is sometimes known as the Chandasekar-Deming estimator after an early application to birth registration completeness (Chandrasekar and Deming 1949). It is often called the dual system estimator (DSE) and can be used to estimate census coverage (see, for example, Marks 1978).

Note that the concept of the DSE rests on the assumption that all units in the population have a nonzero probability of being covered. First, we assume that the event of being included in one system does not change the probability of being included in the other; this is causal independence. Survey management uses administrative steps to try to ensure this assumption holds.

Second, we assume that all units within a frame have the same probability of being included. This probability may differ for each frame, as long as it is constant within a given frame. Since this condition rarely holds, even approximately, for all units in a survey or record system, the DSE is usually calculated on subpopulations where the condition is more likely to hold. For example, while it is unlikely that all farms are included with equal probability, it might be reasonable to assume that all small farms in the South are included with equal probability. In the DSE literature, these separate estimation cells are called poststrata.

The DSE data must be modified to remove duplicates. Essentially, one must remember the numbers in the table above are the counts of true, unique, and correct units included in each system. The literature on the mechanics of measuring coverage using the DSE is large and readily available. See for example, Marks, Seltzer, and Krotki (1974) and Hogan (1992 and 1993).

There are ways to evaluate coverage other than using the DSE. Often post-enumeration surveys try to determine the total population by actually finding all people who were missed in the census. That is, they ignore the *N*[22] cell (i.e., the units not captured in either frame) and instead estimate

$N[**] = N[11] + N[12] + N[21]$

The original Post-Enumeration Survey (PES), conducted in the United States after the 1950 census, used this approach. Several other countries have also used it. This approach seldom works except perhaps when census coverage error is so low that there are few omissions to find. The problem is, of course, that it is even harder to conduct a perfect enumeration (even on a sample basis) several months after the reference date than to do the original count. The second survey often misses more people than the first. Nevertheless, trying to find missed units after the original enumeration can be useful, especially when the characteristics of the missed units can be ascertained (see section 5.4.1).

Post-enumeration surveys have not been limited to human populations. There have been such surveys to measure the coverage of other units, such as schools, housing, and farms. The NCES used a DSE to estimate the 1993–94 Private School Universe. The original frame was constructed using a list frame of private schools. It contained 24,067 schools. An area frame sample was selected to estimate the number of private schools not included on the list frame. This area frame produced an estimate of 21,613 schools of which 19,587 were also found in the original list frame. Thus, the estimated total number of schools would then be

Total = (24,067) * (21,613) / 19,587 = 26,556

Since the private school frame consisted of the two frames combined (area and the list frame), the number of schools using the dual frame estimate is

24,067 + (21,613 - 19,587) = 26,093

Coverage of their combined frame is then estimated as:

Coverage (%) = (26,093 / 26,556) * 100 = 98.3%.

One difficulty of the case-by-case matching approach is that matching errors can and do occur, affecting the accuracy of the measurement of the coverage error.

5.2.3 Other Approaches to Coverage Measurement

A superior estimate may sometimes be constructed on a subsample of cases, where improved, more accurate, and presumably more costly, methods are employed. An example comes from a telephone survey that uses RDD. The RDD excludes households without telephones or with interrupted service. In one type of RDD design, telephone numbers in clusters of telephone numbers ("100-banks") with few residential numbers are excluded. By analyzing results from a personal (face-to-face) interview survey, it is sometimes possible to get information on the characteristics of households missing from the RDD frame. For example, Giesbrecht (1996) and Giesbrecht, Kulp, and Starer (1996) matched telephone numbers collected in the Current Population Survey to the RDD frame, allowing them to determine the number and characteristics of households not covered by various RDD sampling plans.

Lessler and Kalsbeek (1992) and United Nations (1982) also describe reinterview studies (re-enumeration of a sample of the original sample) and record check studies to develop coverage estimates.

5.3 Assessing the Effect of Coverage Errors on Survey Estimates

The literature on coverage measurement focuses on estimating the number of missing or erroneous units and their characteristics. Less work has been devoted to measuring the effect missing units have on survey results, such as, for example, statistics of unemployment or total sales.

Estimating bias resulting from coverage errors is a fairly difficult task. It is unlikely that useful independent sources exist that permit estimation of coverage bias. Consider a survey that measures retail sales by surveying stores during a given month in a given city. If one had a recent economic census list and could identify stores from the sample survey's frame on the census list, then one has both an estimate for the number of missing stores and an estimate for the volume of sales (from the economic census data) missing in the survey data. The percent missing based on the economic census data provides a good estimate for the undercoverage of the sales data. There may be monthly variation in the undercoverage for the monthly survey, but we will not know that. We do, however, have a good estimate for the average annual impact of undercoverage on the sales data.

Matching studies are useful here, but entail their own set of problems. The matching studies will have identified a number (*N*[21]) of units in the population that were not included in the survey frame. However, if the alternate source was an administrative record system or other pre-existing data file, it is not likely to contain the same information the survey would have collected on the missed units. For example, a survey on the health conditions of one-month old babies can be matched against birth records. The birth records of the missing babies may provide useful information about weight at birth, sex, or race, but will not include information about the baby's health at one month. The survey researchers must make use of the available data to estimate the health of the identified missing one-month old babies.

When the second frame is under the direct control of the survey manager it is possible to again ask many or all of the important items. This information allows us to say something about the characteristics of the missed units included in the second frame (*N*[21]). However, it says nothing directly about the units estimated to be missed by both frames (*N*[22]). One way of evaluating the units in *N*[22] is to assume that units missed-by-both frames have the same characteristics as units missed-by-one frame. If a two-way match has been performed, we compare characteristics of the *N*[12] units with those of the *N*[21] units. If these are quite similar, one would be comfortable in assuming that the *N*[22] units are similar as well.

The bias introduced by missing units will depend on several things. Lessler and Kalsbeek (1992) analyze the problem in the following manner:

Define

$\overline{Y_o}$ = Mean of omitted population

$\overline{Y_r}$ = Mean of frame population

N = Number in target population

N_o = Number omitted (=*N*[12])

Let r be the ratio of the population mean of the omitted units $\left(\overline{Y}_0\right)$ to the mean of the frame population ($\overline{Y_r}$).

$$r = \frac{\overline{Y}_0}{\overline{Y}_r}$$

W_o is the proportion of units omitted,

$$W_o = \frac{N_o}{N}$$

The relative bias for estimating a population total is

$$\frac{-W_0 r}{rW_0 + (1 - W_0)}$$

This will be small whenever either r or W_o is small. For example, mobile food carts are often omitted in surveys of retail trade. Although they may represent a measurable proportion of retail outlets W_o, their average sales $\overline{Y_o}$ are much smaller than for shops and stores $\overline{Y}_r$, so r is relatively small. The coverage bias for total sales is considered ignorable.

When estimating means the situation differs. Here the relative bias may be written as

$$\frac{W_0(1 - r)}{(1 - W_0) + rW_0}$$

Obviously, there is no bias if the mean of the omitted units is the same as the mean of the included units (r = 1). As long as r is close to unity and W_o is small, the relative bias on the mean is ignorable.

Indeed to have a large effect on the estimated population means, the population not covered by the survey must be large and quite different from the covered population. For example, assume the survey covers only 90 percent of the population and estimates 5 percent of the population is unemployed, infected, smokes, or has some other characteristic. If the proportion possessing this characteristic among those missed is three times greater (i.e., 15 percent), then the true proportion would be 0.9 (0.05) + 0.1 (0.15) = 0.06 or 6 percent rather than the 5 percent estimated from the frame population.

An example of measuring the effect of coverage error comes from the National Household Education Survey (NHES), a data collection system of the NCES. The NHES is a RDD telephone survey and only includes persons who live in households with telephones. Approximately 6 percent of all persons live in households without telephones, according to data from the March 1992 CPS.[2] The CPS does not systematically exclude nontelephone households. The percentage of persons who live in households without telephones varies by characteristics of the population considered. For example, while 95 percent of all adults live in telephone households, only 87 percent of black adults and 88 percent of Hispanic adults live in telephone households (U.S. Department of Education 1996b).

An important focus of the NHES was on statistics for the population 0- to 2-years old who were sampled as part of the Early Childhood Program Participation (ECPP) component. Supplements

[2] Which, of course, is subject to undercoverage problems of its own.

to the October 1992 CPS were used to examine the extent of the differences in the characteristics of persons in the telephone households and the nontelephone households. The items included in the supplement were limited, containing items about care arrangements and disabilities. More information was gathered on adults.

By tabulating the characteristics of the telephone and nontelephone households from the CPS, an estimate of the bias was made due to excluding nontelephone households in the NHES and ECPP. As a result, some conclusions follow:

> "The analysis of undercoverage bias shows that the coverage biases for estimates of adult characteristics are not very large, while for 0- to 2-year-olds, the biases are somewhat larger, but still relatively small. The undercoverage bias for subgroups…may be more problematic. No specific rule can handle all the subgroups that may be considered by analysts of the NHES:95, but some guidelines are possible. When dealing with a small subgroup that is likely to be differentially covered, analysts need to account for both sampling errors and nonsampling errors. For example, estimates from the NHES for a poorly-covered subgroup such as black children might be approached differently than analysis of all children. Therefore, it is recommended that estimated differences between poorly-covered and well-covered groups (such as black and nonblack children) be considered substantively important only if the differences are larger than both the sampling error and potential coverage bias error (U.S. Department of Education 1996b)."

5.4 Correcting for Coverage Error

There are two general approaches to overcome coverage error. First, improve the frame before data are collected and second adjust the data after they are collected. The most straightforward approach is to take steps to improve the survey frame. Occasionally, this may be done by putting more time, money, or staff into frame development. For example, one might add a quality control step to address listing, or one might work with organizations that have special knowledge about the target population, including professional and trade organizations, or local governments. Acquiring data from administrative and other sources may also improve the frame coverage. Or, one might decide to include telephone banks with only one listed residential number in a survey that had previously excluded these telephone numbers and households. Such improvements are survey specific. They are seldom inexpensive, but nonetheless can prove cost effective (Lessler and Kalsbeek 1992).

5.4.1 Dual Frame Approach

A related approach is to use two or more complimentary sampling frames, often called the dual frame approach. For example, the main sampling frame for the CPS is the list of addresses enumerated in the previous census. This frame is reasonably complete and allows the sample design to use very small sampling clusters, four housing units, at reasonable cost.

Of course, this primary CPS frame excludes all housing units constructed since the previous census. Since one would expect new construction to be closely related to economic growth and the unemployment rate, an aging census frame could contain serious omissions. Therefore, the census list frame is supplemented with a frame based on building permits issued after the census.

Since the frames can be easily unduplicated, they form the basis of improved estimation. Often, a relatively expensive area sample is used to supplement a telephone or list sample. For this approach to work, one must have access to an affordable and accurate way to unduplicate the population covered by each frame.

Consider again the Private School Universe Survey (PSS) discussed above. A list frame was constructed from multiple sources, with the intent to include all private schools. In addition, a complete area frame was constructed by finding and listing all private schools in the sample areas, and an area sample selected. Those private schools already appearing on the list frame were then deleted from the area frame list, and only the previously unlisted schools interviewed. In this way, coverage was improved. In school year 1993–94, the area frame accounted for 7.8 percent of the estimated total number of private schools. The addition was much higher for some private school subgroups. For example, the area sample accounted for 15.3 percent of "unaffiliated" religious schools and 20.5 percent of "special emphasis" schools (U.S. Department of Education 1996a).

5.4.2 Poststratification

In surveys in which there are auxiliary variables that can be used to poststratify, coverage bias may be reduced. The U.S. Bureau of the Census discontinued the area sample in the Monthly Retail Trade Survey because it had access to survey controls from the annual retail trade survey and the Census of Retail Trade. Thus, it was able to use poststratification to help account for missing retail establishments. Poststratification is defined as a process by which all units in the sample are classified into groups or estimation cells. This classification usually takes place after sample selection and data collection.

For each poststratum, the per-unit value is estimated. For example, the sample cases in a demographic survey might be poststratified into male or female; and black, white, or Hispanic. The unemployment rate (for example) for each poststratum could be computed. Because of coverage errors, some groups (e.g., black males) may be underrepresented in our sample. However, if the proportion of each poststratum in the population as a whole is known, the estimated unemployment rate for each poststratum can be multiplied by the population proportions to produce more accurate national estimates of unemployment. The estimate is then said to have been corrected, adjusted, or controlled to population totals. The population information is usually described as population controls or control totals. Often these controls are based on a recent census.

Poststratification, obviously, works well when the noncovered population is similar to the covered population in the post-stratum. Thus to be effective, the poststratification variables must be correlated with the variables of interest. They must also be well measured in the survey and the control totals must be available for the population as a whole. Race and sex are obviously correlated with unemployment, but so is geography, age, etc.

Because of these complexities, survey results are sometimes controlled in several different dimensions in a process known as raking or iterative proportional fitting. The survey totals may first be forced to agree with population estimates by race, and then by sex, etc. For example, monthly CPS estimates of the number of persons in households by race, sex, marital status of householders are used as control totals for the Survey of Income and Program Participation (U.S. Bureau of the Census 1998). Since it operates on survey totals, post-stratification can simultaneously control for coverage, nonresponse, and sampling errors. Clearly, poststratification can be somewhat of a “fix.” This is why coverage ratios should be reported for the survey results before poststratification.

5.5 Reporting Coverage Error in Federal Surveys

During the last 10 years, coverage as a source of error in censuses and surveys has received considerable attention. The U.S. Decennial Census and its undercount of minorities has heightened awareness of this source of error. The FCSM working papers on the quality of establishment data (Federal Committee on Statistical Methodology 1988) and on survey coverage itself (Federal Committee on Statistical Methodology 1990) have described and discussed the issue in some detail and have contributed to the awareness. Despite the continuing and strong interest in this important topic, the reporting of coverage as a source of error remains inconsistent and incomplete.

In their review of analytic publications, Atkinson, Schwanz, and Sieber (1999) report that only 49 percent of the publications they studied specifically mentioned coverage error as a possible source of nonsampling error, and only 16 percent provided an estimated coverage rate. Information about the universe and the sampling frame were more commonly reported.

The cost and complexity of measuring coverage error results in substantial difficulty in reporting quantitative evidence on this source of survey error. Typically, coverage studies are reported as a technical report or a special study where detailed tables provide estimates of undercounts on many characteristics, such as in the 1992 Census of Agriculture (U.S. Bureau of the Census 1996). Regrettably, these studies are reported late after the initial results are released, and, therefore, are ignored by policymakers who use the survey data; however, these studies can identify changes that need to be made to the survey, and thus the studies can help improve the quality of the data.

User’s Guides and Quality Profiles, where substantial information can be summarized, are designed to help the user analyze and understand the data’s limitations. By design, these types of reports are the best vehicles for communicating such information. The Quality Profile for the Survey of Income and Program Participation (U. S. Bureau of the Census 1998) provides useful summaries of the differential undercoverage of demographic subgroups in the SIPP. Ultimately, electronic linkage of the detailed technical information to the analytic reports will become routine (Giesbrecht et al. 1999).

The nature of the publication and survey (one-time versus continuing) plays a significant role in determining what and how much an analyst reports about this source of error. For analytic reports, the subcommittee recommends the following areas and topics be reported:

- Coverage error should be mentioned explicitly as a source of nonsampling error.

- The target population and frame population should be defined and a clear statement made about exclusions in the frame population. The estimated percent of the excluded frame population should be provided.

- The sampling frame should be identified and described. Information about the frame should be reported, such as the year the frame was developed, whether the frame has changed over time, whether it has been updated for births, deaths, and other relevant changes to the study population, and whether gaps or other problems in the frame exist that would affect its quality.

- If available, an overall coverage rate should be defined and provided to the user.

- References to studies about the sampling frame, its quality, and issues related to coverage should be reported.

- If known, the effect of coverage error on key survey estimates should be reported.

The subcommittee recommends more detailed reporting on this source of error in technical reports and user's manuals, including the following topics:

- A general assessment of the quality of the sampling frame should be provided to the data user. This should include a description and discussion of the limitations of the frame.

- Procedures used to update the frame should be described.

- The excluded survey population should be characterized by the available variables.

- Subpopulation coverage rates should be reported, particularly if the subpopulations are important analytic domains.

- Poststratification procedures and the effects of using such procedures should be described.

- A summary of results from studies that aim to measure coverage error should be provided.

References

Atkinson, D., Schwanz, D., and Sieber, W.K. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: Federal Committee on Statistical Methodology, U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Chandrasekar, C. and Deming, W.E. 1949. "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association.* 44: 101–115.

Cox, B.G., Elliehausen, G.E., and Wolken, J.D. 1989. "The National Survey of Small Business Finances: Description and Preliminary Evaluation." Washington, DC: Finance and Economics Discussion Series, Division of Research and Statistics Division of Monetary Affairs, Federal Reserve Board.

Federal Committee on Statistical Methodology. 1990. *Survey Coverage.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 17).

Federal Committee on Statistical Methodology. 1988. *Quality in Establishment Surveys.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 15).

Giesbrecht, L.G. 1996. *Coverage Bias in Various List-Assisted RDD Sample Designs*. Paper presented at the meeting of the American Association of Public Opinion Research.

Giesbrecht, L.G., Kulp, D. W., and Starer, A. W. 1996. "Estimating Coverage Bias in RDD Samples with Current Population Survey (CPS) Data." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 503–508.

Giesbrecht, L., Miller, R. Moriarity, C., and Ware-Martin, A. 1999. "Reporting Data Quality Information on the Internet." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: Federal Committee on Statistical Methodology, U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 342–354.

Hogan, H. 1993. "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association.* 88: 1,047–1,060.

Hogan, H. 1992. "The 1990 Post-Enumeration Survey: An Overview." *The American Statistician*. 46: 261–69.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Lessler, J. and Kalsbeek, R. 1992. *Nonsampling Errors in Surveys.* New York: John Wiley & Sons.

Marks, E.S. 1978. "The Role of Dual System Estimation in Census Evaluation." In K.J. Krotki (ed.), *Developments in Dual System Estimation of Population Size and Growth.* Edmonton: University of Alberta Press.

Marks, E.S., Seltzer, W., and Krotki, K.J. 1974. *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: The Population Council.

Meier. F. and Moore, T. 1999. "The Effect of Screening on Coverage in the National Health Interview Survey." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 566–569.

Peterson, C.G.J. 1896. "The Yearly Immigration of Young Plaice into the Limfjord from the German Sea." *Report of the Danish Biological Station to the Ministry of Fisheries*. 6: 1–48.

Sarndal, C., Swensson, B., and Wretman, J. 1992. *Model-Assisted Survey Sampling*. New York: Springer-Verlag.

Shryock, H.S., Siegel, J.S., and Associates. 1975. *The Materials and Methods of Demography*. Washington, DC: U.S. Bureau of the Census (Third Printing (rev.)).

United Nations. 1982. *National Household Survey Capability Programme, Non-Sampling Errors in Household Surveys: Sources, Assessment and Control*. New York: United Nations Department of Technical Cooperation for Development and Statistical Office (Preliminary version).

U.S. Bureau of the Census. 1998. *Survey of Income and Program Participation* (SIPP) *Quality Profile* (3rd edition). Washington, DC: U.S. Department of Commerce.

U.S. Bureau of the Census. 1996. *1992 Census of Agriculture: Part 2, Coverage Evaluation.* Washington, DC: U.S. Department of Commerce (AC92-S-2).

U.S. Department of Education. 1996a. *Private School Universe Survey, 1993–94.* Washington, DC: National Center for Education Statistics (NCES 96–143).

U.S. Department of Education. 1996b. *Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95).* Washington, DC: National Center for Education Statistics (Working Paper No. 96–29).

Wolter, K. M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association.* 81: 338–346.

Wright, T. and Tsao, H.J. 1983. "A Frame on Frames: An Annotated Bibliography." In T. Wright (ed.), *Statistical Methods and the Improvement of Data Quality.* New York: Academic Press. 25–72.

Chapter 6

Measurement Error

6.1 Introduction

Measurement error is related to the observation of the variables being measured in a survey, and is sometimes referred to as "observation error." A substantial literature exists on measurement error; see O'Muircheartaigh (1997) for a history of measurement error in surveys and Biemer et al. (1991) for a review of important measurement error issues. Measurement error occurs as part of data collection, as opposed to sampling, nonresponse, coverage, or data processing. It may arise from four sources: the questionnaire, the data collection method, the interviewer, and the respondent.

Measurement error can be characterized as the difference between the value of a variable provided by the respondent and the true (but unknown) value of that variable. The total survey error of a statistic with measurement error has both fixed errors (bias) and variable errors (variance) over repeated trials of the survey. *Measurement bias* or *response bias* reflects a systematic pattern or direction in the difference between the respondents' answers to a question and the correct answer; for example, respondents may tend to forget to report a certain type of income such as interest, resulting in reported income lower than the actual income. S*imple response variance* reflects the random variation in the respondent's answer to a survey question over repeated questioning (i.e., respondents may provide different answers to the same question if they are asked the question several times). Interviewers can be a source of this type of variable error. *Interviewer variance,* the variable effects interviewers have on the respondents' answers, is one form of *correlated response variance,* a component of total survey error that occurs because response errors might be correlated for sample units interviewed by the same interviewer.

One approach to estimating measurement error is to compare the responses received from a survey respondent for specific questions against measures of the same variable from an independent source. As a simple example, if respondents were asked their age, responses could be verified against birth records. However, this true value can be elusive. Even in the simple example of verifying age, one cannot assume for certain that birth records are without errors. Nonetheless, we seek to assess the measurement error present in the survey measures by comparing them to measures from an independent and reasonably valid source. Another approach frequently used involves comparing responses from an original interview to those obtained in a second interview conducted soon after the original interview.

Measurement error comes from four primary sources (Biemer et al. 1991). These are:

- *Questionnaire*: The effect of its design, content and wording;
- *Data Collection Method*: The effect of the mode (e.g., mail, telephone, or in person) of administration of the questionnaire. Respondents may answer questions differently in the presence of an interviewer, over the phone, on the computer, or by themselves;
- *Interviewer*: For a survey that relies on an interviewer to administer, the effect the interviewer has on the response to a question. This may include error the interviewer

introduces by not reading the items as intended or adding other information that may misdirect the respondent. Interviewers may introduce these errors due to inadequate training or inadequate skills; and

- *Respondent*: The effect of the respondents. Respondents, because of their different experiences, knowledge, and attitudes may interpret the meaning of questionnaire items differently.

These four sources are the elements that comprise data collection. The questionnaire is the presentation of the request for information. The data collection mode is how the questionnaire is delivered or presented (self-administered, telephone or in person). The interviewer, in the case of telephone or in-person mode, is the deliverer of the questionnaire. The respondent is the recipient of the request for information. Each can introduce error into the measurement process.

While we generally address these sources separately, they can also interact. For example, interviewers' and respondents' characteristics may interact to introduce errors that would not be evident from either source alone. The sections that follow describe in more detail how each of these sources of errors affect data quality and methods for assessing and reducing their effect.

6.2 Sources of Measurement Error

6.2.1 Questionnaire Effects

The questionnaire is designed to communicate with the respondent in an unambiguous manner. It represents the survey designer's request for information. A substantial literature exists on questionnaire effects; for more information, see Groves (1989), Biemer et al. (1991), and Lyberg et al. (1997). In this section, we list some of the ways in which the questionnaire can introduce error into the data collection process.

Most of the effects described below can be evaluated by randomly assigning sample units to one of two (or more) groups. Each group would receive a different version of the questionnaire. Questionnaires to be compared may differ in question wording, question order, response categories, and so on. If an independent data source were available, then results from the two questionnaire versions could be compared to the external data source to determine the "best" version. Otherwise, the result from the two groups could be compared to each other to determine the extent of any differences in reporting. As another variation, the same group of respondents can be asked similar versions of the same questions at a different point of time, but the questions asked must be those for which answers are expected to remain the same over time.

Specification problems

At the survey planning stage, error can occur because the data specification is inadequate and/or inconsistent with what the survey requires. Specification problems can occur due to poorly worded questionnaires and survey instructions, or may occur due to the difficulty of measuring the desired concept. These problems exist because of inadequate specifications of uses and needs, concepts, and individual data elements. A discussion of specification error can be found in *Statistical Policy Working Paper 15: Quality in Establishment Surveys* (Federal Committee on Statistical Methodology 1988).

Question wording

The questionnaire designer attempts to carefully word questions so he/she will communicate unambiguously. The designer wants the respondent to interpret the question as the designer would interpret the question. Words, phrases, and items used in questionnaires are subject to the same likelihood of misunderstanding as any form of communication. The potentials for error are many. First, the questionnaire designer may not have a clear formulation of the concept he/she is trying to measure. Next, even if he/she has a clear concept, it may not be clearly represented in the question. And, even if the concept is clear and faithfully reproduced, the respondent may not interpret the request as intended. Not all respondents will understand the request for information, due to language or cultural differences, affective response to the wording, or differences in experience and context between the questionnaire author and the respondent.

Question comprehension involves at least two levels of processes. One is simply understanding the literal meaning of a sentence. Does the respondent know the words included in the sentence? Can the respondent recall information that matches his/her understanding of those words and provide a meaningful response? However, providing an answer to a question also involves not only an understanding of the literal meaning of the question but also inferring the questioner's intent; that is, to answer the question, the respondent must determine the pragmatic meaning of the question (Schwarz, Groves, and Schuman 1995). It is this second element that makes questionnaire wording development more than just constructing items that have a low enough reading level. Getting feedback from respondents as to how they interpret the intention of items is key to a well-designed instrument. This is typically accomplished through the use of cognitive research methods (see section 6.3.2).

Length of the questions

The questionnaire designer is faced with the dilemma of keeping questions short and simple while assuring sufficient information is provided to respondents so they are able to answer a question accurately and completely. Common sense and good writing practice tell us that keeping questions short and simple will lead to clear interpretation. Research, however, suggests that longer questions actually yield more accurate detail from respondents than shorter questions, at least as they relate to behavioral reports of symptoms and doctors visits (Marquis and Cannell 1971) and questions on alcohol and drug use (Bradburn, Sudman, and Associates 1979). Longer questions may provide more information or cues to help the respondent remember and more time to think about the information being requested.

The effect of question length may be measured if an independent source of data is available by randomly assigning sample units to one of two groups, one receiving a "short" version of the questions and the other group receiving the "long" version of the questions. Responses for each group can then be compared with the "known" values for these questions.

Length of the questionnaire

Long questionnaires may introduce error due to respondent fatigue or loss of concentration. Length of the questionnaire may also be related to nonresponse error, discussed in chapter 4. There is always a tension between the desire to ask as many questions as possible and the awareness that error may be introduced if the questionnaire is too long. The point at which a respondent's attention will be lost will vary depending on respondent characteristics, salience of the topic to the respondent, the interviewer's rapport with the respondent, design of the questionnaire, and mode of interview. If an independent data source is available, the impact of questionnaire length may be tested using a designed experiment. In this experiment, the questions are split into two halves. The question sets appear in reverse order on the two questionnaires.

Order of questions

Asking questions may in and of itself affect how respondents answer later questions, especially in attitude and opinion surveys, where researchers have observed effects of the question order (Schuman and Presser 1981). Assimilation, where subsequent responses are in the same direction as preceding items, and contrast, where subsequent responses are in the opposite direction as preceding items, have been observed. Respondents may also use information from previous items about what selected terms mean to help answer subsequent items. The effect of question order can be assessed by administering alternate forms of a questionnaire to random samples.

Response categories

Response categories help the respondent decide what is important in a question. The respondent infers that categories included with an item are considered the most important by the questionnaire developer. If the respondent does not see the categories he/she believes are appropriate, he/she may become confused as to the intent of the question.

The order of the categories may also affect responses. Response tendencies may incline respondents to typically respond at the same point on a response scale, respond to earlier choices rather than later ones, or choose the later responses offered.

The effect of the order of the response categories may be influenced by the mode of administration. If items are self-administered, response categories earlier in the list are more likely to be recalled and agreed with (primacy effect), because there is more time for the respondent to process them. If items are interviewer-administered, the latter categories are more likely to be recalled (recency effect). Similarly to assessing question order effect, the order of response options can be assessed by testing differing response orders with randomized designs.

Open and closed formats

Question formats in which respondents are asked to respond using a specified set of options (closed format) may yield different responses than when respondents are not given categories (open format) (Bishop et al. 1988). A given response is less likely to be volunteered by a respondent in an open format than when included as an option in a closed format (Bradburn 1983; Molenaar 1982). The closed format may remind respondents of something they may not have otherwise remembered to include. The response options to a question cue the respondent as to the level or type of responses considered appropriate. See, for example, Schwarz, Groves, and Schuman (1995) and Schwarz and Hippler (1991).

Questionnaire format

For the self-administered questionnaire the design and layout of the instrument may help or hinder accurate response. The threat is that a poor design may confuse respondents, lead to a misunderstanding of skip patterns, fatigue respondents, or contribute to their misinterpretation of questions and instructions. Jenkins and Dillman (1997) provide principles for designing self-administered questionnaires. Cognitive research methods provide information to asses the design and format of questionnaires (see section 6.3.2).

6.2.2 Data Collection Mode Effects

Various methods or modes are available for collecting data for a survey. The selection of the data collection mode is a complex decision that depends on the methodological goals of the survey as well as consideration of various factors such as funds available, the questionnaire content, the population covered, expected response rates, length of the collection period, and expected level of measurement error. Lyberg and Kasprzyk (1991) present an overview of different data collection methods along with the sources of measurement error for these methods. A summary of this overview is presented below.

Face-to-face interviewing

Face-to-face interviewing is the mode in which an interviewer administers a structured questionnaire to respondents. Using a paper questionnaire, the interviewer completes the questionnaire by asking questions of the respondent. This method, the *paper and pencil personal interview (PAPI)* method, has a long history of use. Although this method is generally expensive it does allow a more complex interview to be conducted. This mode also allows the use of a wide variety of visual aids to help the respondent answer the questions. A skillful interviewer can build rapport and probe for more complete and accurate responses.

The advent of lightweight laptop personal computers has resulted in face-to-face interviewing being conducted via *computer assisted personal interviewing (CAPI)*. Interviewers visit the respondents' homes and conduct interviews using laptop computers rather than paper questionnaires. The use of CAPI permits editing data for accuracy and completeness at the time of the interview and provides for the correct following of skip patterns. A discussion of the issues related to CAPI can be found in Couper et al. (1998).

One problem for face-to-face interviewing is the effect of interviewers on respondents' answers to questions, resulting in increases to the variances of survey estimates (see section 6.2.3). Another possible source of measurement error is the presence of other household members who may affect the respondent's answers. This is especially true for topics viewed as sensitive by the respondents. Measurement error may also occur because respondents are reluctant to report socially undesirable traits or acts. De Maio (1984) notes that social desirability seems to share two elements—the idea that some things are good and others are bad, and the idea that respondents want to appear "good" and answer questions to appear that way.

Telephone interviewing

This mode is very similar to face-to-face interviewing except interviews are conducted over the telephone rather than in person. Telephone interviewing is usually less expensive and interviews often proceed more rapidly. However, this mode also provides less flexibility (e.g., visual aids cannot be used easily, and complicated and open-ended questions are more difficult to administer). Response rates for telephone surveys have been falling in part because of the use of answering machines to screen calls.

This mode can be conducted from the interviewers' homes or from centralized telephone facilities. Centralized telephone interviewing makes it possible to monitor interviewers' performance and provide immediate feedback. In both cases, either paper questionnaires or computerized questionnaires (*Computer Assisted Telephone Interviewing (CATI)*) can be used. The use of CATI permits editing data for accuracy and completeness at the time of the interview.

Since the interviewer plays a central role in telephone interviewing as well, the sources of measurement error are very similar to those in face-to-face interviewing although the anonymity of the interviewer may improve reporting on sensitive topics by providing adequate "distance" between interviewer and respondent. A discussion of the issues related to telephone interviewing can be found in Dillman (1978); Groves et al. (1988); and Groves (1989).

Self-administered mail surveys

In mail surveys, the questionnaires are mailed to the ultimate sampling units (e.g., a household or a business establishment). The respondents complete and mail back the questionnaire. For demographic surveys, response rates tend to be lower in a survey using a self-administered questionnaire, although Dillman (2000, 1991, and 1983) argues that it is possible to overcome historical limitations of this method and increase response rates to acceptable levels. Self-administered mail surveys are the most commonly used data collection mode for economic surveys.

Mail surveys have different sources of measurement error than face-to-face and telephone interviewing. This mode has no interviewer effects and less risk of "social desirability" effects. However, this mode is more susceptible to misreading and misinterpretation of questions and instructions by the respondents. Good questionnaire design and formatting are essential to reduce the possibility of these problems. Much of the research on questionnaire design and formatting has been conducted in the context of household and demographic surveys, but during the last decade there has been more interest in the application of cognitive research methods to business surveys. See, for example, Phipps, Butani, and Chun (1995); Gower and Nargundkar (1991); Willimack, Nichols, and Sudman (1999).

Diary surveys

Diary surveys are usually conducted for topics that require detailed behavior reporting over a period of time (e.g., expenditures, time use, and television viewing). The respondent uses the diary to enter information about events soon after they occur to avoid recall errors. Interviewers are usually needed to contact the respondent to deliver the diary, gain the respondent's cooperation and explain the data recording procedures, and then again to collect the diary and, if it is not completed, to assist the respondent in completing the diary. Because of the need for a high level of commitment, diary reporting periods are fairly short (typically varying in length from 1 day to 2 weeks).

Lyberg and Kasprzyk (1991) identify a number of sources of measurement error for this mode such as, respondents giving insufficient attention to recording events and then failing to record events when fresh in their memories; the structure and complexity of the diary can present significant practical difficulties for the respondent; and respondents may change their behavior as a result of using a diary.

Computer assisted self-interviewing (CASI)

This mode uses computer technology to obtain information from respondents without the use of interviewers. Couper et al. (1998) address a variety of issues concerning computer assisted surveys. There are several variations of this mode and these are discussed below.

Touchtone data entry (TDE) is used for respondents to answer questions using the keypad of their touchtone telephone. This method is suitable for surveys containing a few simple items. Phipps and Tupek (1991) report on an assessment of measurement errors for a TDE system used for a monthly establishment survey. This assessment was based on three data sources for about 465 Pennsylvania business establishments and found few serious problems with TDE.

Voice recognition entry (VRE) is very similar to TDE except that respondents answer questions by speaking directly into the telephone. The computer translates a respondent's answers into text for verification with the respondent. The main limitation is the current state of technology in voice recognition. Byford (1990) identifies three problems that can affect the quality of voice recognition: wide variations in pronunciation, the large vocabulary of possible words that people use, and the ways in which people run words together.

A third example of CASI is called **prepared data entry (PDE).** In this mode, the respondent reads the survey questions from a computer screen and keys in her/his own answers. An interviewer may bring the computer to the respondent's home or the respondent may be invited to a nearby facility equipped with computers. The interviewer may be present to assist at the beginning, but the respondent keys in the answers. Alternative approaches are mailing preprogrammed floppy disks to the respondent or having the respondent access the questionnaire over the Internet. The latter approach has developed rapidly under the generic term of "*web surveys*." Couper (2001) points out there are various types of web surveys using probability-based and nonprobability-based methods and provides a review of the methods and their associated sources of error.

A fourth example is called **Audio Computer-Assisted Self-Interviewing (ACASI)**. In this example, recent computer technology allows conducting self-administered interviews in which the text on the computer screen is accompanied by a high quality voice recording played over headphones. This application has the potential to remove literacy barriers to self-administered surveys and provide privacy for reporting on sensitive subjects (Lessler and O'Reilly 1995).

Direct observation

Direct observation is a method of data collection where the interviewer collects data by direct observation using his/her senses (vision, hearing, touching, testing) or physical measurement devices. This method is used in many disciplines. For example: 'eye estimation' of crop yield may be used in agricultural surveys; and an electronic measuring device may be used to record television viewing in market research.

Measurement errors may be introduced by observers in ways similar to the errors introduced by interviewers; for example, observers may misunderstand concepts and misperceive the information to be recorded, and may change their pattern of recording information over time because of complacency or fatigue. Fecso (1991) describes the use of direct observation and associated measurement error in crop yield surveys.

Mixed data collection mode

Two or more modes of data collection are used for some surveys to save money, improve coverage, improve response rates, or to reduce measurement errors. There are two types of mixed mode applications. One is a mixed mode application that uses two or more sampling frames. For example, this approach is used to improve coverage in a random digit dial (RDD) survey. An area or address-based sample with face-to-face interviewing is combined with an RDD sample and telephone interviewing to improve coverage.

The second type of mixed mode application can occur in the following situations:

- One main data collection mode is used for a survey and a second or third mode is used for nonresponse followup to improve response rates. For example, a mail survey may use telephone interviews and/or face-to-face interviewing for nonresponse followup to improve the overall response rate (Jabine 1994; Kalton et al. 2000).

- One mode is used to screen a population to identify a subpopulation with rare attributes and another mode is used to interview this subpopulation. For example, telephone interviewing may be appropriate to identify the people with the desired characteristics and the followup interview of these people may be conducted by mail or personal interview.

- In panel surveys, a mixed mode strategy is used primarily to reduce costs and take advantage of the best features of each mode. Face-to-face interviewing in the first round of a panel survey improves on the incomplete coverage associated with nontelephone households and also establishes a relationship with the sample unit to help maintain cooperation and response in future rounds of the panel (Kalton, Kasprzyk, and McMillen 1989).

Comparison of mode effects

Research into effects of data collection modes on data quality have generally focused on the three modes that are used most frequently. They are face-to-face interview, the telephone interview, and the self-administered (mail) questionnaire. The National Center for Health Statistics (NCHS) conducted a formal randomized experiment in which random half-samples of households were assigned either to telephone or face-to-face interviews. The telephone treatment was then randomly split into a computer assisted telephone interview (CATI) treatment and a PAPI telephone interview. As such, researchers could compare face-to-face interviewing to telephone interviewing (combined CATI and paper-and-pencil), to CATI, and to paper-and-pencil interviewing. The CATI and PAPI techniques could also be compared. Thornberry (1987) presents findings of this research concluding that initial concerns about “major differences in data quality between the ongoing National Health Interview Survey (NHIS) and a telephone NHIS were largely unfounded.” See Sudman and Bradburn (1974); Lyberg and Kasprzyk (1991); deLeeuw (1993); and deLeeuw and Collins (1997) for additional examples. Nicholls, Baker, and Martin (1997) describes new technologies and summarize research studies on the effects of collection technologies on survey data quality.

6.2.3 Interviewer Effects

Because of individual differences, each interviewer handles the survey situation in a different way, that is, in asking questions, probing and recording answers, or interacting with the respondent, some interviewers appear to obtain different responses from others. The interviewer situation is dynamic and relies on an interviewer establishing rapport with the respondent. Interviewers may not ask questions exactly as worded, follow skip patterns correctly or probe for answers nondirectively. They may not follow directions exactly, either purposefully or because those directions have not been made clear enough. Interviewers may vary their inflection, tone of voice, or other personal mannerisms without even knowing it. Errors, both overreports and underreports, can occur for each interviewer. When overreporting and underreporting of approximately the same magnitude occurs, small interviewer bias will result. However, these individual interviewer errors may be large and in the same direction, resulting in large errors for individual interviewers. To the extent these errors are large and systematic, a bias, as measured in the mean squared error of the estimate, will result and this is called the interviewer effect. Another potential source of interviewer effects is respondent reaction to characteristics of the interviewer, such as age, race, sex, or to attitudes or expectations of the interviewer.

Correlated interviewer variance

In the early 1960's attention turned to estimating the size of the interviewer effect and three different approaches were suggested (Hansen, Hurwitz, and Bershad [1961]; Kish [1962]; and Fellegi [1964]). Hansen, Hurwitz, and Bershad; and Kish presented an intra-interviewer correlation coefficient. Fellegi expanded on the intra-interviewer correlation coefficient developed by Hansen, Hurwitz, and Bershad and presented a more complex model involving reinterviewing to estimate this component. The Kish approach is the least demanding in terms of experimental design and permits comparisons between studies involving different numbers of interviewers and respondents. It measures the interviewer effect in terms of *rho*, the intra-interviewer correlation coefficient, defined as the ratio of the interviewer variance component to the total variance of a survey variable and estimated by a simple analysis of variance.

In well-conducted face-to-face surveys the *rho* typically clusters around 0.02 and in controlled telephone surveys the value of *rho* averages below 0.01 (Hox, deLeeuw, and Kreft 1991). Although these proportions are small, the effect on the precision of the estimate may be large. The variance of the sample mean is multiplied by 1 + *rho*(n-1), where *n* is the size of the average interviewer workload. A *rho* of 0.02 with a workload of 10 interviews increases the variance 18 percent. A workload of 25 yields a variance 48 percent larger. Thus, even apparently small interviewer intraclass correlations can produce important losses in the precision of survey statistics. For practical and economic considerations, each interviewer usually has a large workload. An interviewer who is contributing a systematic bias will thus affect the results obtained from several respondents and the effect on the variance is large.

Interviewer characteristics

There is no basis for recommending that recruitment of interviewers should be concentrated among women rather than men, or among middle class persons, or among the middle-aged rather than the young or the old (Collins 1980). However, interviewers may contribute to errors in estimates through their complex personal interactions with respondents. Weiss (1968) studied a sample of welfare mothers in New York City, validated the accuracy of several items on the survey, and found that similarity between interviewer and respondent with respect to age, education, and socioeconomic status was not necessarily conducive to more valid reporting. There are instances, however, in which better reporting occurred when the interviewer had a higher level of education than the respondent. Groves and Fultz (1985) studied interviewer gender differences as a source of error and found that on a variety of factual items no effects of interviewer gender were observed, although on attitudinal items greater optimism was expressed when the interviewer was male. Groves (1989) reviewed a number of studies and concluded, in general, demographic effects appear to apply when the measurements are related to the characteristics but not otherwise. That is, there may be an effect based on the race of the interviewer if the questions asked were related to race.

Other interviewer factors may also play a role in interviewer-produced error, such as voice characteristics and interviewing expectations. Oksenberg and Cannell (1988) studied vocal characteristics of telephone interviewers and their relation to refusal rates. Sudman et al. (1977) studied interviewer expectations about the difficulty of obtaining sensitive information and observed weak effects on the relationship between expectations of difficulties in interviewing and actual difficulties encountered.

Methods to control interviewer errors

Three different means to control interviewer errors are: training, supervision or monitoring, and workload manipulation. Many believe that standardization of the measurement process especially as it relates to interviewers' tasks leads to a decrease in interviewer effects. One way to accomplish standardization is through a training program of sufficient length, usually 2.5 to 3 days, to cover interview skills and techniques as well as information on the specific survey (Fowler 1991).

Supervision and performance monitoring are essential ingredients of a quality control system. The objectives of such a system are to monitor interviewer performance through observation and performance statistics and identify problem questions. Reinterview programs and field observations are conducted to evaluate individual interviewer performance. Observations in the field are conducted using extensive coding lists or detailed observers' guides where the supervisor or monitor checks whether the procedures are properly followed. For instance, the observation could include the interviewer's grooming and conduct, introduction of himself/herself and of the survey, the manner in which the questions are asked and answers recorded, the use of flash cards and neutral probes, and the proper use of the interviewers' manual. In other instances, tapes (either audiovisual or audio) can be made and interviewer behavior coded (Lyberg and Kasprzyk 1991).

A third way to control interviewer effects, at least from a bias point-of-view, is to change the average workload; however, as mentioned above, interviewer variance increases as average workload increases. The issue is to find the optimal average workload. Groves and Magilavy (1986) discuss optimal workload as a function of interviewer hiring and training costs, interview costs, and size of intra-interviewer correlation. Determining optimal workload is a difficult task because the value of the intra- interviewer correlation varies among statistics in the same survey.

Interviewer effects can be reduced by avoiding pitfalls of questionnaire design, giving clear and unambiguous instructions and definitions, training interviewers to follow these instructions and by minimizing the reliance on the variable skill of interviewers to extract information.

6.2.4 Respondent Effects

Respondents may contribute to error in measurement by failing to provide accurate responses. Groves (1989) indicates that both traditional models of the interview process (Kahn and Cannell 1957) and the cognitive science perspectives on survey response (Hastie and Carlston 1980) identify the following five sequential stages in the formation and provision of answers by survey respondents:

- *Encoding of information*: involves the process of forming memories or retaining knowledge;

- *Comprehension of the survey question*: involves knowledge of the words and phrases used for the question as well as the respondent's impression of the purpose of the survey, the context and form of the question, and the interviewer's behavior in asking the questionl;

- *Retrieval of information from memory:* involves the respondent's attempt to search her/his memory for relevant information;

- *Judgment of appropriate answer:* involves the respondent's choosing from the alternative responses to a question based on the information that was retrieved; and

- *Communication of the response*: involves the consideration of influences on accurate reporting that occur after the respondent has retrieved the relevant information as well as the respondent's ability to articulate the response.

There are many aspects of the survey process that can affect the quality of the respondent's answers resulting from this five-stage process. Examples of factors that can influence respondent effects follow.

Respondent rules

One survey factor related to the response process is the respondent rules (i.e., the eligibility criteria used for identifying the person(s) to answer the questionnaire). For surveys collecting information for the sample unit (e.g., households, businesses, schools), the specific respondent's knowledge about the answers to the questions may vary among the different eligible respondents. Respondent rules in establishment surveys, for example, can be complicated depending on the nature of the data request. Responses to a survey about school districts are dependent on whether the questions asked concern policy of the district about an issue in education, school district finances, or school district enrollment and staffing. Similarly, data requests made of large companies may not be completed by a single respondent or office. Nichols, Willimack, and Sudman (1999) identify three groups of data provider for establishment surveys: 1) those at the corporate office; 2) those at business unit level; and 3) those at the establishment level or individual business level.

Surveys collecting information for individuals within the sample unit (e.g., persons within households, employees within businesses, and students and teachers within schools) may use self-reporting or proxy reporting. Self versus proxy reporting differences vary by subject matter (e.g., self-reporting is generally better for attitudinal surveys). Mathiowetz and Groves (1985) and Thornberry (1987) report on results of randomized experiments that collected health data over the telephone using two different respondent rules (i.e., randomized respondents and knowledgeable adult respondents). Blair, Menon, and Bickart. (1991) present a literature review of self versus proxy reporting research.

Questions

The respondent's comprehension of a question is affected by the wording and complexity of the question, and the design of the questionnaire (see section 6.2.1 for more details). The respondent's ability to recall the correct answer is affected by the type of question asked and by the difficulty of the task in determining the answer. The respondent's willingness to provide the correct answer to questions is affected by the type of question being asked, by the difficulty of the task in determining the answer, and by the respondent's view concerning the social desirability of the responses.

Interviewers

The respondent's comprehension of the question is affected by the interviewer's visual clues (e.g., age, gender, dress, facial expressions) as well as audio cues (e.g., tone of voice, pace, inflection). See section 6.2.3 for more details.

Recall period

The longer the time period between an event and the survey the more likely it is respondents will have difficulty remembering the activity the question is asking about. The converse will be true for a shorter time period. Survey designers need to identify the recall period that minimizes the total mean squared error in terms of the sampling error and possible biases. For example, Huang (1993) presents the results of a recall period study done for the Survey of Income and Program Participation (SIPP) to determine if SIPP should change from a 4-month to a 6-month reference period. This study found that the increase in precision would not compensate for the increase in bias so it was decided to continue the 4-month reference period. Eisenhower, Mathiowetz, and Morganstein. (1991) indicate that memory aids (e.g., calendars, maps, diaries) have been used for reducing recall bias. Mathiowetz (2000) reports the results of a meta analysis that tests the hypothesis that the quality of retrospective reports is a function of the length of recall period. The conclusions depend on whether you examine the direction of the relationship between length of recall period and data quality or the size of the effect. The data suggest the nature of the data request is an important consideration that affects the quality of the data.

Telescoping

Telescoping occurs when respondents report occurrences within the recall period when they actually occurred outside the recall period. Bounding techniques (e.g., conduct an initial interview solely to establish a reference date, or use a significant date or event as the beginning of the reference period) can be used to reduce the effects of telescoping. These techniques were developed by Neter and Waksberg (1964) in a study of recall of consumer expenditures. For example, the National Crime Victimization Survey uses an initial unbounded interview to establish a reference date for future interviews with their sample households.

Panel/Longitudinal surveys

In addition to the other factors previously described, panel or longitudinal surveys that interview the same sample cases several times, have some additional respondent-related factors. First, spurious measures of change over time may occur when a respondent reports different answers at two different points in time due to random variation in answering the same questions rather than due to a real change. The Survey of Income and Program Participation Quality Profile (U.S. Bureau of the Census 1998) discusses this in some detail. Dependent interviewing techniques, in which the responses from the previous interview are used in the data collection, can be used to reduce the incidence of these spurious changes for questions that tend to have unreliable gross change measures when independent interviewing is used. Hill (1994) found that dependent interviewing resulted in a net improvement in measures of change in occupation and industry of employment. Dependent interviewing can result in missing reports of true change, so selectivity in the use of dependent interviewing is necessary. Mathiowetz and McGonagle (2000) review current practices within a computer assisted interviewing environment as well as the empirical evidence with respect to the impact of dependent interviewing on data quality.

Another source of concern in panel surveys is panel conditioning or "time-in-sample" bias. Conditioning refers to a change in response that occurs because the respondent has had one or more prior interviews. For example, in the National Crime Victimization Survey, respondents are interviewed seven times at 6-month intervals. Woltman and Bushery (1977) took advantage of the design of this survey, which represents a randomized experiment with the number of previous interviews as a treatment, to investigate time-in-sample bias for this survey. They compared victimization reports for these treatments with varying degrees of panel experience (i.e., previous interviews) who were interviewed in the same month. They found generally declining rates of reported victimization as the number of previous interviews increased. A further discussion of this phenomenon can be found in Kalton, Kasprzyk, and McMillen (1989).

6.3 Approaches to Quantify Measurement Error

6.3.1 Randomized Experiments

Groves (1989) indicates that randomized experiments are a frequently used method for both estimating measurement errors that vary over replications of a survey as well as those that are fixed features of a design. Survey researchers have given this method a variety of names such as interpenetrated samples, split-sample experiments, split-panel experiments, random half-sample experiments, and split-ballot experiments.

In this technique, random subsamples of identical design are administered different treatments related to the specific error being measured. For studying variable errors, many different entities thought to be the source of the error are included and compared (e.g., many different interviewers for interviewer variance estimates). For studying biases, usually only two or three alternative designs (treatments) are compared (e.g., two different data collection modes) with one of the methods being the preferred method. Often randomized experiments to evaluate alternative methods, procedures, and questionnaires are included in field tests conducted prior to the fielding of a survey. These experiments are used to identify approaches that minimize measurement error.

6.3.2 Cognitive Research Methods

To design questionnaires that are as free of measurement error as possible, survey designers turn to respondents to help them. No matter how skilled or experienced questionnaire designers are, they cannot be sure that the respondents will interpret the items as they intend them to. Thus, they can always benefit by testing questionnaires with respondents similar to those responding when the questionnaire is fielded. This type of testing is called cognitive testing. It differs from field testing or psychometric testing in that the respondents provide information to the designer on how they interpret the items in the questionnaire. Because it is labor intensive and therefore costly per respondent, cognitive testing is normally done on samples that are small relative to those used for field tests.

Respondents are asked to complete the draft questionnaire and to describe how they interpret each item. An interviewer will probe regarding particular words, definitions, skip patterns, or other elements of the questionnaire on which they wish to get specific feedback from the respondent. Respondents will be asked to identify anything not clear to them. Respondents may be asked to do this as they complete the questionnaire or in a debriefing session afterwards. The process is somewhat interactive in that the designer may add probes and pursue the clarity of different items or elements of the questionnaire in subsequent interviews. This ongoing feedback process lends itself to an iterative approach.

In situations in which respondents provide information on how they interpret survey items as they are filling out the questionnaire, they may be asked to “think aloud,” thus verbalizing what they are thinking about as they respond to the items. The advantage of this technique is that it can yield information the designers would not anticipate. It also is not subject to interviewer-imposed bias. The disadvantage is that it does not work well for respondents who are not comfortable with or skilled at verbalizing their thoughts in this way (Willis 1994).

An associated technique involves having an interviewer ask the respondent about some feature of the question immediately after the respondent completes an item (Nolin and Chandler 1996). This approach is less dependent on the respondent’s comfort and skill level with verbalizing his/her thoughts. It, however, limits the investigation of the items to those things the survey designer can think to ask about. Also the approach may introduce interviewers’ bias since they administer the probes. Some questionnaire designers consider the probing approach artificial since it is different from the usual interview situation (Willis 1994). These approaches, of course, can be combined and the interviewer may probe areas the respondent does not cover in “thinking aloud” or areas that need clarification.

Other approaches allow the respondent to complete the survey instrument before being asked to provide feedback on their interpretation of the items. The interviewer may ask the respondent to think aloud about the questionnaire items or the interviewer may ask specific questions that the questionnaire designer is interested in. Delayed questioning may be done one-on-one or in focus groups. While some respondents may be less comfortable than others, focus groups provide the advantage of the interaction of group members which may lead to focus groups exploring areas that might not be touched on in one-on-one interviews. A weakness of cognitive interviews is that they are done with small nonrandom sample of cases. The designer does not know whether the information obtained from respondents is representative of the potential respondents for the survey.

Expert panels, a small group of experts brought in to critique a questionnaire, can be an effective way to identify problems in the questionnaire (Czaja and Blair 1996). Survey design professionals and/or subject-matter professionals receive the questionnaire several days prior to a meeting with the questionnaire designers. In a group session, the individuals review and comment on the questionnaire on a question by question basis.

Methodological issues in the application of cognitive psychology to survey research are discussed by Tucker (1997). Cognitive research methods have become critical to the questionnaire design process and the reduction of measurement error in surveys. Sudman, Bradburn, and Schwarz (1996) present a summary of major findings as they relate to survey methodology.

6.3.3 Reinterview Studies

Reinterview—a replicated measurement on the same unit in interview surveys—is a new interview which reasks the questions of the original interview (or a subset of them) for a small subsample (usually around 5 percent) of a survey's sample units. It is conducted for one or more of the following four purposes:

- To identify interviewers who are falsifying data;
- To identify interviewers who misunderstand procedures and require remedial training;
- To estimate simple response variance; and
- To estimate response bias.

The first two purposes provide information on measurement errors resulting from interviewer effects. The last two purposes provide information on measurement errors resulting from the joint effect of all four sources (i.e., interviewer, questionnaire, respondent, and data collection mode). Reinterviews do not usually provide an estimate of response variance/bias attributed to each source.

The remaining sections describe the specific design requirements for each of these four types of reinterviews, as prescribed by Forsman and Schreiner (1991). In addition, some methods for analyzing these reinterviews along with limitations of the results are also presented.

Interviewer falsification reinterview

The intentional falsification of survey results by interviewers can take a variety of forms (e.g., an interviewer can make up answers for some or all of the questions, or an interviewer can deliberately deviate from stated survey procedures). A reinterview sample is drawn and is generally completed by telephone, whenever possible, to reduce travel costs for the supervisory staff. A falsification rate, defined as the proportion of interviewers falsifying interviews detected through the falsification reinterview, can be derived. In table 6.1, Schreiner, Pennie, and Newbrough. (1988) provide the following rates of interviewer falsification for three surveys conducted by the U.S. Bureau of the Census:

Table 6.1.—Rates of interviewer falsification, by survey

Survey	Rate of interviewer falsification
Current Population Survey	0.4%
National Crime Victimization Survey	0.4%
New York City Housing and Vacancy Survey	6.5%

SOURCE: Schreiner, I., Pennie, K., and Newbrough, J. 1988. “Interviewer Falsification in Census Bureau Surveys.” *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 491–496.

Interviewer evaluation reinterview

Interviewer evaluation reinterviews identify interviewers who misunderstand survey procedures and target them for additional training. Most design features for this type of reinterview are identical to those for a falsification reinterview. The results of the interviewer evaluation reinterview are used to determine whether interviewer performance is acceptable. Tolerance tables, based on statistical quality control theory, are used to determine if the number of differences in the reinterview after reconciliation exceed a specific acceptable limit. Reinterview programs at the U.S. Bureau of the Census use acceptable quality tolerance levels ranging between 6 and 10 percent (Forsman and Schreiner 1991).

Simple response variance reinterview

The simple response variance reinterview must be conducted as an independent replication of the original interview procedures. Thus, the reinterview subsample should be a representative subsample of the original sample design. The pool of interviewers used for the initial interviews should conduct the reinterview. The reinterview respondents should be selected using the respondent rules applied in the original interview. If possible, the original interview questionnaire should be used for the reinterview. If this questionnaire is too long, a subset of the original interview questionnaire is used. Differences between the original interview and the reinterview should *not* be reconciled. The data collection mode used for the original interview should also be used for the reinterview.

One common statistic that is estimated from a simple response variance reinterview is the *gross difference rate (GDR)*, which is the average squared difference between the original interview and reinterview responses. The GDR divided by 2 is an unbiased estimate of simple response variance (SRV). For characteristics that have two possible outcomes, the GDR is equal to the percentage of cases that had different responses in the original interview and the reinterview. Brick, Rizzo, and Wernimont. (1997) provide general rules to interpret the response variance measured by the gross difference rate.

Another common statistic is the *index of inconsistency (IOI)*, which represents the proportion of the total population variance caused by simple response variance.

$$IOI = \frac{GDR}{s_1^2 + s_2^2}$$

Where s_1^2 and s_2^2 are the sample variances for the original interview and the reinterview respectively.

A common interpretation of the IOI values is as follows:

- An IOI of less than 20 is *low* relative response variance;
- An IOI between 20 and 50 is *moderate* relative response variance; and
- An IOI above 50 is *high* relative response variance.

The GDR and IOI response variance measures from reinterviews provide data users with information on the reliability and response consistency for a survey's questions as well as identify those questions that are problematic.

Feindt, Schreiner, and Bushery (1997) describe a survey program's efforts to continuously improve questionnaires. In this process, cognitive research and other questionnaire design methods are conducted on questions having higher rates of discrepancy as identified in the reinterview. These methods may help determine the cause of the problems and identify improvements that can be made to solve these problems. In the next round of survey interviews, a reinterview is conducted on these revised questions to determine whether reliability improvements have been made. This process is then repeated for the remaining problematic questions. Feindt, Schreiner, and Bushery (1997) give results of this process for the National Center for Education Statistics' Schools and Staffing Survey (see table 6.2).

Table 6.2.—Summary of the reinterview reliability for the 1988 and 1991 administrator and teacher surveys

Which of the following degrees have you earned?	Index of inconsistency			Gross difference rate (%)		
	1988	1991	Z(diff)	1988	1991	Z(diff)
Teacher						
Bachelor's degree	79.5	—	([2])	7.5	0.6	[1]6.8
Master's degree	8.9	2.2	[1]3.8	4.3	1.1	[1]3.7
Administrator						
Bachelor's degree	98.5	—	([2])	20.3	1.3	[1]14.5
Master's degree	49.4	11.3	[1]6.7	9.9	1.7	[1]7.4

— Too few cases to reliably estimate the index.
[1] significant at 0.10 alpha level.

[2] Not applicable.

SOURCE: Feindt, P., Schreiner, I., and Bushery, J. 1997. "Reinterview: A Tool for Survey Quality Management." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 105–110.

Response bias reinterview

The response bias reinterview is conducted in a manner that produces the "true" or correct responses to the extent possible. The reinterview subsample must be a representative subsample of the original sample design. Usually the most experienced interviewers and supervisors are used to conduct the reinterview. The most knowledgeable respondent is used as the reinterview respondent or each person answers questions for themselves, so the use of proxy respondents is minimized. The original interview questions can be used for the reinterview, and the differences between the two responses are reconciled with the respondent to establish "truth." It is helpful that the questions used be expected to remain constant over time. An alternative approach is to use a series of probing questions to replace the original questions in an effort to obtain the true responses and then to reconcile differences with the respondent. Face-to-face interviews should be used for this type of reinterview.

The use of reconciliation to establish truth does have limitations. The respondents may knowingly report false information and consistently report this information in the original interview and the reinterview so that the reconciled reinterview will not yield the "true" estimates. In a study of the quality of the Current Population Survey reinterview data, Biemer and Forsman (1992) determined that up to 50 percent of the errors in the original interview were not detected in the reconciled reinterview.

Response bias is estimated from the response bias reinterview by calculating the *net difference rate (NDR)*, which is the average difference between the original interview response and the reconciled reinterview response assumed to represent the "true" answer.

$$NDR = \frac{1}{n} \sum_{i=1}^{n} (y_{Oi} - y_{Ti})$$

Where: *n* is the reinterview sample size;

y_O is the original interview response; and

y_T is the reinterview response after reconciliation which is assumed to be the true response.

The NDR response bias measures from reinterviews provide data users with information about the accuracy of a survey question and also identify those questions providing biased results. The existence of this bias needs to be considered when these data are analyzed and the results interpreted. An example of the use of reinterview to estimate response bias can be found in Brick et al. (1996). In this example an intensive reinterview was conducted with a sample of respondents and responses were compared to original values. The intensive reinterview was used to obtain more detailed and accurate information by better understanding the respondent's perspective and reason for his/her answers. Although working with a small sample, the authors concluded that the method has potential for detecting and measuring biases. Bias-corrected estimates were developed, illustrating the potential effects on estimates when measures of bias are available.

6.3.4 Behavior Coding

Behavior coding is a technique for collecting data to evaluate interviewer performance. The system uses codes which encompass all of the interviewer's major verbal activities and is designed for use in both training and on-the-job supervision. For each interviewer behavior, such as question asking, probe usage, response summarization, and other behavior of the interviewer, codes are assigned to record interviewer's actions. For example, several codes classify the interviewer's reading of a question: there is a code for questions which he/she asks correctly and completely, one for those which he/she asks with minor changes and omissions, and one for those which he/she either rewords substantially or does not complete.

Overall, the coding system indicates whether questions were asked correctly or incorrectly, whether probes directed the respondent to a particular response or further defined the question or were non-directive, whether responses were summarized accurately or inaccurately, and whether various other behaviors were appropriate or inappropriate. The coded results reflect the degree to which the interviewer employs the methods in which he/she has been trained. That is, an "incorrect" or "inappropriate" behavior is defined as one which the interviewer has been trained to avoid. To establish and maintain a high level of coding reliability for each coded interview, a second coder should independently code a subsample of interviews.

A behavior coding system is useful in three ways: 1) In initial training, it teaches the novice interviewer which interviewing techniques are acceptable and which are not; 2) It serves as a basis for interviewers and supervisors to review work in the field and discuss the problems which coding reveals; and 3) It provides an assessment of an interviewer's performance, which can be compared both with the performance of other interviewers and with the individual's own performance during interviews (Cannell, Lawson, and Hauser 1975).

Oksenberg, Cannell, and Blixt (1996) describe a study in which interviewer behavior was tape recorded, coded, and analyzed for the purpose of identifying interviewer and respondent problems in the 1987 National Medical Expenditure Survey (NMES) conducted by the Agency for Health Care Policy and Research (now called the Agency for Healthcare Research and Quality). The main objective of the study was to see how interview behavior deviated from the principles and techniques covered in the interviewers' training. A major finding was that interviewers frequently deviated from asking the questions as worded, at times in ways in which they could influence responses. In addition, interviewers did not probe as much as necessary, but when they did, the probes tended to be directive or inappropriate.

6.3.5 Interviewer Variance Studies

Interviewer variance studies are studies in which statistical modeling is used to obtain a measure of interviewer effects. Because these models assume randomization, some method of randomizing the interviewers with the respondents is needed so that the differences in the results obtained by different interviewers can then be attributed to effects of the interviewers themselves. In designs with "replication" two interviewers are randomly assigned to the same respondent. Comparisons of the results are used to measure the impact that the interviewer has on the survey responses. In this design the two answers to a single question might be related to one another because the interviewers share some attribute or because the respondents remember their first response. For this reason, differences between responses obtained by the interviewer and reinterviewer do not only measure interviewer variance but also effects of memory of the first response.

"Interpenetrated" interviewer assignments avoid multiple interviews with the same respondent. They estimate interviewer variance by assigning each interviewer to different but similar respondents, that is, respondents who have the same attributes on the survey variables. In practice, this equivalency is assured through randomization. That is, the sample is partitioned into subsets at random, each having the same attributes in expectation as the others and then each interviewer works on a different subset. With this design each interviewer conducts a small survey with all the essential attributes of the large survey except its size.

Interpenetrated interviewer assignments take a different form in personal interview surveys compared to centralized telephone surveys. In personal interview survey designs, interviewer assignments are geographically defined to avoid large traveling costs. The assigned areas have sizes sufficient for one interviewer's workload. Pairs of assignment areas are identified and randomly assigned to pairs of interviewers. Within each assignment area each interviewer of the pair is assigned a random half of the sample housing units. Thus, each interviewer completes interviews in two assignment areas and each assignment area is handled by two different interviewers. The design consists of one experiment (a comparison of results of two interviewers in each of two assignment areas) replicated as many times as there are pairs of interviewers. Bailey, Moore, and Bailar (1978) present an example of interpenetration for personal interviews in the National Crime Victimization Survey (NCVS) in eight central cities. More recently, O'Muircheartaigh and Campanelli (1998) showed interview variance can be as large as the sampling variance due to the geographic clustering of households in postal code sectors.

In centralized telephone surveys, there are no cost savings enjoyed by restricting the randomization of cases to geographical areas. Furthermore, most facilities allow assignment of any sample case to any interviewer working in the facility. The entire sample that is active at any one shift can be partitioned into one of six priority groups (with 1 meaning "must call" and 6 meaning "must not call"). The units within each priority group to be called (1 through 5) are randomly assigned to the interviewers, but if one priority group is exhausted before units from the next priority are assigned, then each priority-shift level plays the role of the enumeration area pair in the personal survey. In this case, each shift is treated as a new group and comparisons of results obtained by different interviewers are conducted within shifts.

6.3.6 Record Check Studies

A record check study involves a comparison of survey results for individual sample cases with an external source generally assumed to contain the true value for the survey variables. Such studies are used to estimate response bias resulting from the joint effect of all four sources of measurement error (i.e., interviewer, questionnaire, respondent, and data collection mode). These studies do not usually provide an estimate of response bias attributed to each source.

Groves (1989) describes the following kinds of record check study designs:

- The reverse record check study;
- The forward record check study; and
- The full design record check study.

In a *reverse record check study*, the survey sample is selected from a source that contains accurate data on the important characteristics under study. The response bias estimate is then based on a comparison of the survey responses with record data.

Reverse record check studies cannot measure errors of overreporting (i.e., falsely reporting an event). These studies can only measure what portion of the sample source records contain events reported in the survey and whether the characteristics of these events on the sample source records are the same as those from the survey. For example, a reverse record check study was conducted by the Law Enforcement Assistance Administration (1972) to assess errors in reported victimization. Police department records were sampled and the victim in the record was contacted. Table 6.3 illustrates the results of the study, indicating that 74 percent of the crimes were reported by the victims during the interview.

Table 6.3.—Cases sampled from police records by whether crime was reported in the survey "within past 12 months" by type of crime

Type of crime	Total police cases interviewed	Percentage reported to interviewer as "within past 12 months"
All crimes	**394**	**74.1**
Violent crimes	206	62.6
Assault	81	48.1
Rape	45	66.7
Robbery	80	76.3
Property crimes	188	86.2
Burglary	104	90.3
Larceny	84	81.0

SOURCE: Law Enforcement Assistance Administration. 1972. *San Jose Methods Test of Known Crime Victims*. Washington, DC (Statistics Technical Report No.1). 6, Table C.

In a *forward record check study*, external record systems containing relevant and accurate information on the survey respondents are located after the survey responses are obtained. Response bias estimates are then based on a comparison of the survey responses to the records from these external systems for the characteristics contained on both the survey and the external sources. The strength of forward record check studies is the ability to measure overreporting. However, they do require contacting several different record keeping agencies and obtaining permission from the respondents to obtain this information. These studies are also limited in measuring underreporting.

Chaney (1994) describes a forward record check study for comparing teachers' self-reports of their academic qualifications with college transcripts. Transcripts were requested for interviewed teachers who were asked to provide a list of all colleges attended, and a total of 1,524 transcripts were received for these teachers. Table 6.4 shows the results of comparing the teachers' self-reports on the year they earned their academic degrees with these transcripts.

Table 6.4.—Accuracy of teachers' self-reports on the year they earned their academic degrees

Comparison of self-reports and transcript data	Bachelor's degree		Master's degree		Associate's degree		Doctoral degree	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total	**427**	**100**	**137**	**100**	**19**	**100**	**4**	**100**
Transcript matches self-report	374	88	99	72	13	68	4	100
Transcript conflicts with self-report	53	12	39	28	6	32	0	0
Direction of discrepancy								
Self-report too recent	22	5	19	14	2	11	0	0
Self-report too early	31	7	20	15	4	21	0	0
Size of discrepancy								
1 year	28	7	27	20	5	26	0	0
More than 1 year	25	6	12	9	1	5	0	0

NOTE: Percentages may not sum to totals because of rounding. Cases with missing data on the year the degree was earned are excluded.

SOURCE: Chaney, B. 1994. *The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (Working Paper No. 94–04).

In a *full design record check study*, features of both the reverse and forward record check designs are combined in that a sample is selected from a frame covering the entire population and records from all sources relevant to the sample cases are located. As a result, errors associated with underreporting and overreporting can be measured by comparing the survey responses to all records (i.e., from the sample frame as well as from external sources) for the survey respondents. Although this type of record check study avoids the weakness of the reverse and forward record check studies, it does require a database that covers all units in the population and all the corresponding events for those units. Marquis and Moore (1989a, 1989b, and 1990) and Marquis, Moore, and Huggins (1990) provide a detailed description and the design and analysis of a full record check study conducted to estimate measurement errors in the Survey of Income and Program Participation (SIPP). In this study, survey data on program and benefit amounts for eight Federal and State benefit programs in four states were matched against the administrative records for the same programs. The *SIPP Quality Profile* (U.S. Bureau of the Census 1998) provides a summary of the design and analysis. For this study, the U.S. Bureau of the Census obtained complete program data files for the states and time periods covered by the study and conducted the linkages using a combination of computer and manual matching procedures. The results provide estimates of the percent response bias in SIPP estimates of the level of participation in the eight programs (see table 6.5).

Table 6.5.—SIPP record check: Percent net bias in estimates of program participation

Program	Percent bias*
Social security retirement	1
Veterans' benefits	-3
Civil service retirement	-8
Supplement security income	-12
Food Stamps	-13
Workers' compensation	-18
Unemployment insurance	-20
AFDC	-39

* Negative bias indicates net underreporting.

SOURCE: Adapted from Marquis, K.H., Moore, J.C., and Huggins, V.J. 1990. "Implications of SIPP Record Check Results for Measurement Principles and Practice." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 564–569.

All three types of record check studies share three other limitations. First, it is assumed that the record systems are free of errors of coverage, nonresponse, or missing data. Second, it is also assumed that the individual records on these systems are complete, accurate, and free of measurement errors. The third limitation involves matching errors—errors that occur as part of the process of matching the respondents' survey records with their administrative records.

For all three types of record check studies, an estimate of response bias can be obtained for a given characteristic by estimating the average difference between the survey response and the record check value for that characteristic. The following formula can be used:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)$$

Where: n is the record check study sample size;

Y_i = survey response for the i^{th} sample person; and

X_i = record check value for the i^{th} sample person.

The response bias measures from a record check study provide data collectors and data users with information about the accuracy of a survey question and identify those questions that are producing biased results. Response bias measures can also be used to make improved estimates for these questions (see Salvucci et al. 1997).

Response bias measures from a record check study can also be used for evaluating alternatives for various survey design features such as questionnaire design, recall periods, data collection modes, and bounding techniques. For example, the results of a reverse record check study in three counties of North Carolina regarding motor vehicle accident reporting is presented in Cash and Moss (1972). Interviews were conducted in 86 percent of the households containing sample persons identified as involved in motor vehicle accidents in the 12-month period prior to the interview. Table 6.6 shows higher proportions of persons not reporting the accident when interviewed many months after the accident. Only 3.4 percent of the accidents occurring within 3 months of the interview were not reported, but over 27 percent of those occurring between 9 and 12 months before the interview were not reported.

Table 6.6.—Percentage of respondents not reporting the motor vehicle accident, by number of months between accident and the interview

Number of months	Percentage not reported	Number of persons
Less than 3 months	3.4	119
3–6 months	10.5	209
6–9 months	14.3	119
9–12 months	27.3	143

SOURCE: Cash, W.S. and Moss, A.J. 1972. *Optimum Recall for Reporting Persons Injured in Motor Vehicle Accidents.* National Center for Health Statistics. 2(50), Table C.

6.4 Reporting Measurement Error in Federal Surveys

Measurement errors may arise in respondents' answers to survey questions for a variety of reasons, including misunderstanding the meaning of the question, failure to recall the information correctly, failure to construct the response correctly (e.g., summing the components of an amount incorrectly). In personal interview surveys, measurement errors may arise from interviewers who may cause the respondents to provide inaccurate responses by asking the question or probing incorrectly, misinterpreting responses, or making errors in recording responses.

Measurement errors are the most difficult errors to quantify. Special studies using randomization of subsamples, reinterviews, record checks, and approaches discussed earlier, are necessary, and they are usually expensive to conduct. Nevertheless, if specific measurement issues are obvious during the planning of the survey, implementing the means to quantify the key error source is desirable. The analysis of measurement error studies typically lags behind the release of the data and are reported in methodological reports, at professional meetings, or in peer-reviewed journals. In many cases, this lag is substantial, and thus an understanding of the limitations of the data is rarely appreciated in time to have an impact on analysis. Nevertheless, it is important for analytic publications to acknowledge the existence of this error and when possible refer the reader to completed, ongoing, or related studies. In regularly conducted survey programs, it is advantageous for studies to be brought together in synthesis reports or quality profiles to help the data user get an understanding of the quality of the survey data.

In the absence of complex and costly studies to quantify aspects of measurement error, indicators of the steps taken to reduce or minimize measurement are useful, if only to indicate the quality of the survey operations to the end user of the data. Background information about the planning of the survey, pretests, small-scale experiments, cognitive research on the questionnaire, interviewer training, and other survey research activities designed to reduce measurement error should be reported to help the data user judge the overall quality of the data collection program.

Reported improvements in data collection and statistical procedures implemented to reduce errors in repeated surveys can serve as implicit data quality indicators.

In their study of a limited number of analytic publications, Atkinson et al. (1999) found measurement error specifically mentioned in 67 percent of the publications reviewed. Specific sources were identified and described in 51 percent of the reviewed publications. These reporting rates are higher than anticipated—and that is good—but, in most cases, this literally meant a statement was made to the effect that error from a particular source was possible. Studies designed to quantify this error source were mentioned in 18 percent of the publications. This is not necessarily surprising, although a number of the reviewed publications used data from periodic or continuing surveys where the opportunity for this type of error measurement presents itself regularly.

Because of the difficulty, complexity, and expense in quantifying measurement error, the publications reviewed did not report very much detail on this particular error source. However, a data user cannot understand the limitations of the data—from a measurement error point of view—unless the data collection program takes steps to explicitly provide such information. The studies required are costly, time consuming, and not available quickly. Recommendations for reporting this source error must take the practical realities into consideration. Thus, for analytic reports, the subcommittee recommends:

- Measurement error, in general, ought to be defined and described as a source of nonsampling error.

- Examples of different sources of measurement error likely to be found in the survey should be given.

- Studies, such as reinterviews, record check studies, or split-sample experiments, to quantify and understand measurement error in the context of the survey ought to be briefly summarized, if available, with references made to detailed methodological or technical reports.

- The amount of information and detail reported on sources of measurement error is related to the relative importance of the source of error, what is known about the source of error, and how it may affect characteristics analyzed in the report.

- Implicit data quality indicators, such as steps taken to reduce measurement error (for example, pretests, experiments, interviewer training, and cognitive testing of questionnaires) should be reported.

- References should be provided to synthesis reports or quality profiles that describe the variety of measurement studies conducted, their results and the possible effects on analysis.

- In general, the amount of detail reported concerning sources of measurement error should be dependent on the known or assumed effects of the source of error on key statistics.

Technical reports, user's manuals, and quality profiles are the appropriate dissemination venues for detailed reporting on the planning, conduct, and analysis of measurement error studies. These detailed reports are most useful to the data user if the results can be related to the key statistics and findings of the analytic report.

References

Atkinson, D., Schwanz, D., and W.K. Sieber. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Bailey, L., Moore, T.F., and Bailar, B.A. 1978. "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample." *Journal of the American Statistical Association.* 73: 16–23.

Biemer, P.P. and Forsman, G. 1992. "On the Quality of Reinterview Data with Application to the Current Population Survey." *Journal of the American Statistical Association*. 87: 915–923.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (eds.) 1991. *Measurement Errors in Surveys*. New York: John Wiley & Sons.

Bishop, G.F., Hippler, H.J., Schwartz, N., and Strack, F. 1988. "A Comparison of Response Effects in Self-administered and Telephone Surveys." In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg, (eds.), *Telephone Survey Methodology.* New York: John Wiley & Sons. 321–340.

Blair, J., Menon, G., Bickart, B. 1991. "Measurement Effects in Self vs. Proxy Responses to Survey Questions: An Information-Processing Perspective." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*. New York: John Wiley & Sons. 145–166.

Bradburn, N.M. 1983. "Response Effects." In P.H. Rossi, J.D. Wright, and A.B. Anderson (eds.), *Handbook of Survey Research.* New York: Academic Press. 289–328.

Bradburn, N.M., Sudman, S., and Associates. 1979. *Improving Interviewing Methods and Questionnaire Design: Response Effects to Threatening Questions in Survey Research.* San Francisco: Jossey-Bass.

Brick, J.M., Rizzo, L., and Wernimont, J. 1997. *Reinterview Results for the School Safety and Discipline and School Readiness Components.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 97–339).

Brick, J.M., Kim, K., Nolin, M.J., and Collins, M. 1996. *Estimation of Response Bias in the NHES:95 Adult Education Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (Working Paper No. 96–13).

Byford, R.G. 1990. "Advances in Voice Data Collection." *Transactions of 44th Annual Quality Congress.* American Society for Quality Control. 592–600.

Cannell, C.F., Lawson, S.A., and Hauser, D.L. 1975. *A Technique for Evaluating Interviewer Performance*. Ann Arbor, MI: The University of Michigan. 3–95.

Cash, W.S. and Moss, A.J. 1972. "Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents." *Vital and Health Statistics.* Washington, DC: Public Health Service. 2(50).

Chaney, B. 1994. *The Accuracy of Teachers' Self-reports on Their Post Secondary Education: Teacher Transcript Study, Schools and Staffing Survey*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (Working Paper 94–04).

Collins, M. 1980. "Interviewer Variability: A Review of the Problem." *Journal of the Market Research Society*. 22(2): 77–95.

Couper, M.P. 2001. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly*. 64(4): 464–494.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L., and O'Reilly, J.M. (eds.). 1998. *Computer Assisted Survey Information Collection.* New York: John Wiley & Sons.

Czaja R. and Blair, J. 1996. *Designing Surveys: A Guide to Decisions and Procedures*. Thousand Oaks, CA: Pine Forge Press, A Sage Publications Company.

deLeeuw, E.D. 1993. *Mode Effects in Survey Research. A Comparison of Mail Telephone and Face to Face Surveys*. BMS 41, 3–14.

deLeeuw, E.D. and Collins, M. 1997. "Data Collection Methods and Survey Quality: An Overview." In L. Lyberg, P. Biemer, M. Collins, E.D. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons: 199–220.

De Maio, T.J. 1984. "Social Desirability and Survey Measurement: A Review." In C.F. Turner and E. Martin (eds), *Surveying Subjective Phenomena.* New York: Russell Sage. 257–282.

Dillman, D.A. 2000. *Mail and Internet Surveys: The Tailored Design Method.* New York: John Wiley & Sons.

Dillman, D.A. 1991. "The Design and Administration of Mail Surveys." *Annual Review of Sociology.* 17: 225–249.

Dillman, D.A. 1983. "Mail and Other Self-administered Questionnaires." In P. Rossi, R.A. Wright, and B.A. Anderson (eds.), *Handbook of Survey Research.* New York: Academic Press. 359–377.

Dillman, D.A. 1978. *Mail and Telephone Surveys: The Total Design Method.* New York: John Wiley & Sons.

Eisenhower, D., Mathiowetz, N.A., and Morganstein, D. 1991. "Recall Error: Sources and Bias Reduction Techniques." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 127–144.

Fecso, R. 1991. “A Review of Errors of Direct Observation in Crop Yield Surveys.” In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 327–346.

Federal Committee on Statistical Methodology. 1988. *Quality in Establishment Surveys*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 15).

Feindt, P., Schreiner, I., and Bushery, J. 1997. “Reinterview: A Tool for Survey Quality Management.” *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 105–110.

Fellegi, I.P. 1964. “Response Variance and Its Estimation.” *Journal of the American Statistical Association.* 59: 1,016–1,041.

Forsman, G. and Schreiner, I. 1991. “The Design and Analysis of Reinterview: An Overview.” In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 279–302.

Fowler, F.J. 1991. “Reducing Interviewer-related Error Through Interviewer Training, Supervision and Other Means.” In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 259–275.

Gower, A. and Nargundkar, M.S. 1991. “Cognitive Aspects of Questionnaire Design: Business Surveys Versus Household Surveys.” *Proceedings of the 1991 Annual Research Conference*. Washington, DC: U.S. Bureau of the Census. 299–312.

Groves, R.M. 1989. *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L., and Waksberg, J. (eds.). 1988. *Telephone Survey Methodology*. New York: John Wiley & Sons.

Groves, R.M. and Fultz, N. 1985. “Gender Effects Among Telephone Interviewers in a Survey of Economic Attributes.” *Sociological Methods and Research*. 14(1): 31–52.

Groves, R.M. and Magilavy, L.J. 1986. “Measuring and Explaining Interviewer Effects.” *Public Opinion Quarterly.* 50: 251–256.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. 1961. “Measurement Errors in Censuses and Surveys.” *Bulletin of the International Statistics Institute.* 38(2): 359–374.

Hastie, R. and Carlston, D. 1980. “Theoretical Issues in Person Memory.” In R. Hastie et al. (eds.), *Person Memory: The Cognitive Basis of Social Perception*. Hillsdale, NJ, Lawrence Erlbaum. 1–53.

Hill, D.H. 1994. “The Relative Empirical Validity of Dependent and Independent Data Collection in a Panel Survey.” *Journal of Official Statistics.* 10(4): 359–380.

Hox, J.J., deLeeuw, E.D., and Kreft, I.G.G. 1991. "The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 439–461.

Huang, H. 1993. *Report on SIPP Recall Length Study*. Internal U.S. Census Bureau Report.

Jabine, T. 1994. *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 94-340).

Jenkins, C. and Dillman, D. 1997. "Towards a Theory of Self-Administered Questionnaire Design." In L. Lyberg, P. Biemer, M. Collins, E.D. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: JohnWiley & Sons. 165–196.

Kahn, R.L. and Cannell, C.F. 1957. *The Dynamics of Interviewing.* New York: John Wiley & Sons.

Kalton, G., Winglee, M., Krawchuk, S., and Levine, D. 2000. *Quality Profile for SASS: Rounds 1–3, 1987–1995*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 2000–308).

Kalton, G., Kasprzyk, D., and McMillen, D.B. 1989. "Nonsampling Errors in Panel Surveys." In D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh (eds.), *Panel Surveys*. New York: John Wiley & Sons. 249–270.

Kish, L. 1962. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association*. 57. 92–115.

Law Enforcement Assistance Administration. 1972. *San Jose Methods Test of Known Crime Victims.* Washington, DC (Statistics Technical Report No.1).

Lessler, J. and O'Reilly, J. 1995. "Literacy Limitations and Solutions for Self-Administered Questionnaires." *Seminar on New Directions in Statistical Methodology*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 23). 453–469.

Lyberg, L. and Kasprzyk, D. 1991. "Data Collection Methods and Measurement Errors: An Overview." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 237–258.

Lyberg, L., Biemer, P., Collins, M., deLeeuw, E.D., Dippo, C., Schwartz, N., and Trewin, D. 1997. *Survey Measurement and Process Quality.* New York: John Wiley & Sons.

Marquis, K.H. and Cannell, C.F. 1971. "Effects of Some Experimental Techniques on Reporting in the Health Interview." *Vital and Health Statistics.* Washington, DC: Public Health Service. 2(41).

Marquis, K.H. and Moore, J.C. 1990. "Measurement Errors in SIPP Program Reports." *Proceedings of the Bureau of the Census' 1990 Annual Research Conference.* 721–745.

Marquis, K.H. and Moore, J.C. 1989a. "Response Errors in SIPP: Preliminary Results." *Proceedings of the Bureau of the Census Fifth Annual Research Conference*. 515–536.

Marquis, K.H. and Moore, J.C. 1989b. "Some Response Errors in SIPP—With Thoughts About Their Effects and Remedies." *Proceedings of the Section on Survey Research Method.* Alexandria, VA: American Statistical Association. 381–386.

Marquis, K.H., Moore, J.C., and Huggins, V.J. 1990. "Implications of SIPP Record Check Results for Measurement Principles and Practice." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. 564–569.

Mathiowetz, N. 2000. "The Effect of Length of Recall on the Quality of Survey Data." *Proceedings of the 4th International Conference on Methodological Issues in Official Statistics*. Stockholm: Statistics Sweden. Available at http://www.scb.se/omscb/proceedings.asp

Mathiowetz, N. and McGonagle, K. 2000. "An Assessment of the Current State of Dependent Interviewing in Household Surveys." *Journal of Official Statistics*. 16:401–418.

Mathiowetz, N.A. and Groves, R.M. 1985. "The Effects of Respondent Rules on Health Survey Reports." *American Journal of Public Health.* 75(6): 639–644.

Molenaar, N.J. 1982. "Response Effects of Formal Characteristics of Questions." In W. Dijkstra and J. vanderZouwen (eds.), *Response Behavior in the Survey Interview.* New York: Academic Press.

Neter, J. and Waksberg, J. 1964. "A Study of Response Errors in Expenditure Data from Household Interviews." *Journal of the American Statistical Association.* 59: 18–55.

Nicholls, W.L., Baker, R.P, and Martin, J. 1997. "The Effect of New Data Collection Technologies on Survey Data Quality." In L. Lyberg, P. Biemer, M. Collins, E.D. deLeeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality.* New York: John Wiley & Sons. 221–248.

Nichols, E., Willimack, D., and Sudman, S. (1999). "Who are the Reporters: A Study of Government Data Providers in Large Multi-unit Companies." Paper presented at the Joint Statistical Meetings of the American Statistical Association. Baltimore, MD.

Nolin, M.J. and Chandler, K. 1996. *Use of Cognitive Laboratories and Recorded Interviews in the National Household Education Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 96–332).

Oksenberg, L. and Cannell, C. 1988. "Effects of Interviewer Vocal Characteristics on Nonresponse." In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II., and J. Waksberg (eds.), *Telephone Survey Methodology.* New York: Wiley & Sons. 257–269.

Oksenberg, L., Cannell, C., and Blixt, S. 1996. "Analysis of Interviewer and Respondent Behavior in the Household Survey." *National Medical Expenditures Survey Methods 7.* Rockville, MD: Agency for Health Care and Policy Research, Public Health Services.

O'Muircheartaigh, C. (1997). "Measurement Error in Surveys: A Historical Perspective." In L. Lyberg, P. Biemer, M. Collins, E.D. deLeeuw, C. Dippo, N., Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality.* New York: John Wiley & Sons. 29–46.

O'Muircheartaigh, C. and Campanelli, P. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistical Society*, Series A, 161(I). 63–77.

Phipps, P.A., Butani, S.J., and Chun, Y.I. 1995. "Research on Establishment Survey Questionnaire Design." *Journal of Business and Economic Statistics.* 337–346.

Phipps, P.A. and Tupek, A.R. 1991. "Assessing Measurement Errors in a Touchtone Recognition Survey." *Survey Methodology*. 17(1): 15–26.

Salvucci, S., Walter, E., Conley, V., Fink, S., and Saba, M. 1997. *Measurement Error Studies at the National Center for Education Statistics.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 97–464).

Schreiner, I., Pennie, K., and Newbrough, J. 1988. "Interviewer Falsification in Census Bureau Surveys." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 491–496.

Schuman, H. and Presser, S. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Schwarz, N., Groves, R.M., and Schuman, H. 1995. "Survey Methods" *Survey Methodology Program Working Paper Series.* Ann Arbor, MI: Institute for Survey Research, University of Michigan.

Schwarz, N. and Hippler, H. 1991. "Response Alternatives: The Impact of Their Choice and Presentation Order." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys.* New York: John Wiley & Sons. 41–56.

Sudman, S. and Bradburn, N. 1974. *Response Effects in Surveys: A Review and Synthesis.* Chicago, IL: Aldine.

Sudman, S., Bradburn, N., Blair, E., and Stocking, C. 1977. "Modest Expectations: The Effect of Interviewers' Prior Expectations on Response." *Sociological Methods and Research.* 6(2): 171–182.

Sudman, S., Bradburn, N., and Schwarz, N. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Thornberry, O. 1987. "An Experimental Comparison of Telephone and Personal Health Interview Surveys." *Vital Health Statistics*. Washington, DC: Public Health Service. 2(106). DHHS Pub No. (PHS)87–1380.

Tucker, C. 1997. "Methodological Issues Surrounding the Application of Cognitive Psychology in Survey Research." *Bulletin of Sociological Methodology*. 55: 67–92.

U.S. Bureau of the Census. 1998. *Survey of Income and Program Participation (SIPP) Quality Profile*. 3rd edition. Washington, DC: U.S. Department of Commerce.

Weiss, C. 1968. "Validity of Welfare Mothers' Interview Response." *Public Opinion Quarterly*. 32: 622–633.

Willimack, D., Nichols, E., and Sudman, S. 1999. "Understanding the Questionnaire in Business Surveys." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 889–894.

Willis, G.B. 1994. *Cognitive Interviewing and Questionnaire Design; A Training Manual*. Hyattsville, MD: National Center for Health Statistics (Cognitive Methods Staff Working Paper No. 7*)*.

Woltman, H.F. and Bushery, J.B. 1977. "Update of the National Crime Survey Panel Bias Study." Internal U.S. Census Bureau Report.

Chapter 7

Processing Error

7.1 Introduction

A survey consists of many processing steps, from data capture to final publication of survey results. Each step can generate errors in the data or the published statistics. These errors, referred to collectively as processing errors, come in a variety of types that range from simple recording errors (e.g., transcribing or transmission error) to more complex errors arising from mis-specification of an edit or imputation model. The errors that occur for a particular survey are strongly influenced by survey planning, and to some extent the survey's resources (e.g., staff and budget) and constraints (e.g., elapsed time between data collection and publication). In general, resources and constraints weigh heavily in the data collection mode selected, with each mode resulting in different types of processing errors. Opportunities presented by technology, such as the use of computer-assisted techniques and scanning, can minimize processing errors. However, if not carefully managed these technologies can, themselves, introduce errors as a result of data transfer, transmission, or (in the case of scanning) translation problems.

This chapter describes common types of processing errors that occur at the various stages of a survey. It concentrates on the data entry, coding and editing processes, and the errors associated with them. It discusses review and management of the errors and communication of the extent and effect of the errors to the final data users. Since processing errors are often considered part of the administration or operation of the survey itself, several authors have emphasized the need for process control techniques and continuous quality management (e.g., Morganstein and Marker 1997; Linacre and Trewin 1989; Linacre 1991).

The chapter purposely avoids discussion of programming errors in survey instruments or other "gross" errors in preparing data processing systems. While these types of problems can occur, it is assumed that adequate review processes are implemented to avoid them. Ultimately, while some processing errors are inevitable in any large-scale survey, careful survey planning and monitoring is necessary to minimize their frequency and effect.

7.2 Measuring Processing Error

7.2.1 Data Entry Errors

Data entry errors occur in the process of transferring collected data to an electronic medium. The frequency of these errors varies by the types of information collected (e.g., numeric versus character) and the mode of data collection. For example, with paper and pencil enumeration, survey data are key-entered after the survey interview takes place. Data validation for this mode of data entry is obviously different than it is for computer-assisted interviewing modes, where real-time editing and validation are often built directly into the data collection process. With more modern data collection and capture modes, such as those using computer-assisted and web collection techniques and scanning, data entry errors can be considerably reduced. However, data entry errors occur even with technologically advanced techniques. For example, computer-assisted interviewing techniques may have some level of problematic key entry, since interviewers with little data entry experience now enter information. Dielman and Couper (1995) report keying error rates of 0.99 percent in the computer-assisted personal interviewing environment. Scanning for direct data capture also creates errors since it uses imperfect character recognition algorithms. Therefore, regardless of the mode of collection and technology used, data entry errors of some frequency are likely to occur. For a review of the data capture aspects of post-survey processing, see Lyberg and Kasprzyk (1997).

Whatever data collection method is selected, the collection phase of conducting a survey is typically the most expensive part. By comparison, quality control mechanisms to ensure the quality of the captured data are generally very economical. These should be put in place and monitored to ensure that data entry errors are kept to an absolute minimum.

Key entry errors

Double key entry is an effective technique for minimizing key-entry errors in paper and pencil instruments. With this technique, key-entered data are independently keyed a second time, usually by a different key-entry operator, and the resulting two data sets are compared. Discrepancy reports identifying cases and survey items for which differences exist are then reviewed, and erroneous data are corrected. In some cases only a sample of questionnaires is verified to test the accuracy of the key-entry process and of individual key-entry personnel. While somewhat less expensive, this alternative lacks the data correction capability of the 100 percent verification solution. However, with limited budgets it may be a viable lower-cost alternative, as estimates of keying error rates can be developed from these samples. Variations of these verification strategies are also used. For example, in the American Housing Survey the work of new key entry personnel is 100 percent verified until the error rate is at or below a certain level, at which time only a sample of questionnaires is checked (Chakrabarty and Torres 1996). Reports of keying error rates can provide survey managers useful indicators of the efficiency of key-entry for a particular survey.

In most cases keying is not a very error-prone operation. For instance, the error level in the Fourth Followup Survey of the National Longitudinal Study (as determined from a quality control sample of questionnaires) was about 1.6 percent (Henderson and Allen 1981). Similarly, error rates for individual variables in the 1988 U.S. Survey of Income and Program Participation were about 0.1 percent (Jabine, King, and Petroni 1990). According to a study of the quality of the key entry of long form data in the 1990 U.S. Census, key-entry personnel in the initial stages of production committed keystroke mistakes (or omissions) in 0.62 percent of the fields (U.S. Bureau of the Census 1993).

However, key entry has been shown to be a more problematic operation for other surveys. An example of this is described in the Energy Information Administration’s Residential Energy Consumption Survey Quality Profile (U.S. Energy Information Administration 1996). In 1981, 1982, and 1984, key household survey items from the Residential Energy Consumption Survey (RECS) were 100 percent verified, with the remaining items verified at a sample rate of 25 percent. A review of the changes made during the processing for the 1984 survey showed that keying errors were leading to substantial numbers of computer edit rejects (Jabine 1987). The 193 “key” variables for which data entry was 100 percent verified had an average of only 0.44 changes per variable, but the remaining 369 variables that were only subject to sample verification averaged 4.83 changes per variable. Survey managers determined that the processing costs associated with the additional erroneous data exceeded the savings from sample verification of data entry. Consequently, beginning with the 1987 RECS, all keying has been 100 percent verified.

Electronic scanning

With today's technology, electronic scanning can be used effectively to facilitate the handling of data from paper and pencil questionnaires. Scanning of survey data can be used effectively both with and without electronic optical character recognition (OCR). Statistics Canada successfully used scanners without OCR to produce photographic images of their 1996 Census of Agriculture questionnaires. The scanned questionnaire images were then available to be viewed on specially designed display terminals to assist in resolving individual report problems discovered later in the survey process. The scanning initiative was considered a major success, as it saved many hours of retrieving the paper questionnaires. It is also thought to have improved data quality, as editors were more likely to use the original questionnaires in making data decisions, since they were readily accessible in photographic form (Jones and Green 1998).

Taking the technology one step further, scanning with OCR can serve as a data entry option. This approach eliminates the human errors in data entry, but introduces a different source of processing errors. Scanned and computer interpreted, the data must be reviewed since electronic OCR is less than 100 percent accurate. There are two types of OCR errors: rejection and substitution. Rejects are entries that cannot be scanned. They must be corrected and add no error if corrected properly. Substitutes are entries that are read incorrectly (e.g., "1" instead of "7"). In many contexts these substitutions have minimal impact on survey results, especially with demographic data. However, significant errors of this type can occur in business surveys, especially when the misread is in the left-most digit of a report for a large operation. Blom and Lyberg (1998) discuss the history and development of scanning technology and provide a discussion of "rejects" and "substitutes."

More recently, Statistics Sweden and other survey organizations have experimented with scanning entire survey forms with complete OCR. This procedure eliminates data entry in the usual sense, and allows for both editing and storage as part of the scanning operation. The OCR technology, however, is still relatively young, and the step from scanning in-house produced OCR digits to correctly interpreting information in free-form respondent handwriting is a big one. Reject rates on the survey data on position levels from a Statistics Sweden study ranged from 7 to 35 percent. Not surprisingly, questions requiring less free-form handwriting tend to be much more successfully computer interpreted. The U.S. Bureau of the Census used scanning with OCR technology as a major data entry tool for its population census in 2000. The National Agricultural Statistics Service is planning the same approach for its 2002 Census of Agriculture. Blom and Lyberg (1998) discuss the scanning and OCR experiences of national survey organizations.

Indicators of quality of the scanning process include error frequencies and rates by data item. The scanning software available from some vendors produces this information directly in monitoring its own operation. If the available scanning software for a particular application does not directly provide it, the desired information can be obtained through keying a sample of the forms and comparing the scanned data to the key-entered data.

Computer-assisted data entry

With computer-assisted interviewing techniques, the respondent's data are obtained through a personal or telephone interview or a self-administered electronic questionnaire. These media simplify quality control mechanisms. Mandatory read-back validation is often used to ensure that interviewers have captured the respondent's data accurately. Online edits can also be used to check for internal consistency as well as temporal consistency between current and previous reports for the same respondent. Edits that check ranges for continuous variables, limit responses to specific values for categorical data, and check for logical relationships among two or more variables can also be very useful in avoiding data entry errors. Many editing systems are designed to capture edit failure rates as well as correction rates. Considered together, these two rates provide a measure of the quality of a particular edit flag. If a much larger percentage of records are being flagged than are being corrected, then the edit most likely needs to be revised. If, on the other hand, correction percentages are relatively high, these can provide an indicator of either response or data entry problems.

7.2.2 Pre-Edit Coding Errors

Most surveys require some type of pre-edit coding of the survey returns before they can be further processed in edit, imputation, and summary systems. The required coding is generally of two types—unit and item response coding. The unit response coding assigns and records the status of the interview. It is designed to indicate the response status of the return for a sampled unit so that it can be appropriately handled in subsequent processing. Was a usable response obtained? If so, who responded? Does the reporting unit match the sample unit? Unit response coding is typically done shortly after the survey responses are "checked-in" for paper and pencil responses and within the survey instrument for computer-assisted interviewing. Unit response coding is a critical process in classifying a unit as a respondent or nonrespondent. The classification procedure is defined by a fixed set of criteria. For example, in the NCES' Schools and Staffing Survey Public School data file, a school was classified as a respondent if the school reported the number of students, the number of teachers, as well as 30 percent of the remaining items (Gruber, Rohr, and Fondelier 1996).

Item response coding is the more commonly discussed form of questionnaire coding. This can involve coding an actual response for a survey question into a category. This situation occurs for questions that elicit open-ended responses. For example, sometimes one of the responses in categorical questions is "other-specify," which results in a free-text response. In other surveys, the respondent may be asked by design to provide an open-ended response. This is the case in the industry and occupation coding that is required in many federal surveys. For example, the Survey of Income and Program Participation (U.S. Bureau of the Census 1998) questionnaire solicits verbal descriptions of occupation and industry. The responses are transmitted electronically to the U.S. Bureau of the Census' processing center where codes are assigned. Once codes are assigned, they are all verified through a dependent verification process where the verifiers have access to the coders' entries.

The recoding of open-ended responses into a categorical variable is performed by coders who interpret and catalogue each response. This process can result in error or bias, since different coders are likely to interpret and code some responses differently. Even the same coders may change the way they code as they gain more experience or get bored. Another problem with open-ended responses is that respondents sometimes supply more than one answer to the question. For example, when asking respondents why they choose to shop at a particular store, a respondent may say it was convenient and cheap—two distinct answers. This may occur even if instructions request only the most important reason.

A technique employed to measure and reduce coding errors is the use of multiple coders for the same set of responses. This is analogous to double key entry. Once both sets of coders have completed the coding, a comparison is made between the two sets of data, with any discrepancies resolved by committee or by an expert. This is a fairly expensive option, but it results in a more consistent set of data. To reduce the cost somewhat, a sample of cases might be double coded. Based on this sample, a determination can be made as to whether full double coding is required, and whether it is required for all coders or just those experiencing problems.

An obvious indicator of quality from this activity is the percentage of times the coding differed between the coders for each item. Items with large discrepancy rates should be reviewed to determine the causes. Distributions of the reasons for discrepancies can be captured as another indicator of coding process quality.

Formal studies are conducted occasionally to measure the level of reliability in coding operations for open-ended questions. Kalton and Stowell (1979) conducted an experiment to study reliability levels of professional coders and presented results that confirmed the findings of other studies—that coding can contribute substantially to survey error.

Automated coding can sometimes be used to help reduce coder errors. For example, in computer-assisted interviews, the computer can sometimes assign response coding without any interviewer interaction. Whether this is an option in a particular situation often depends on whether there is adequate information in the record for the computer to make a reliable decision. Even if totally automated coding is not possible, there may be enough information for the computer to flag particular cases as problematic and in need of review. The percent of cases requiring review can be a measure of the quality of the coding process.

Automated coding of open-ended industry and occupation responses in population censuses and other demographic surveys was researched and pilot tested in the 1970s and 1980s. The first major implementation of the Automated Industry and Occupation Coding System (AIOCS) was with the 1990 U.S. Decennial Census. The implementation of AIOCS was based on a predecessor system tested by the U.S. Bureau of the Census in the early 1980's. To research the use of this system, the U.S. Bureau of the Census compared automated industry and occupation coding on the Consumer Expenditure Survey to manual clerical coding of the same records (Appel and Hellerman 1983). A panel of experts adjudicated the differences between the manual coders and the automated results. The authors found that a fully automated system can provide significant benefits in terms of cost, timeliness, and data quality.

In contrast to the traditional quality assurance operations that rely on acceptance sampling guaranteeing a pre-specified average outgoing quality, others, such as Biemer and Caspar (1994), suggest a different approach to improving coding operations based on a process of comparing coding performance with preferred performance, identifying nonconforming observations by type, working with teams to identify reasons for nonconforming observations, and implementing suggested measures for improvement.

7.2.3 Editing Errors

Among the more recognized processing errors in statistical literature are those resulting from editing. An earlier Federal Committee on Statistical Methodology (1990) working paper dealt with data editing in federal agencies and discussed its effect on the survey data in considerable detail. This paper defined editing in the following way:

> "Procedure(s) designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures."

Editing generally occurs at various points in the survey processing and can, in itself, generate errors at each juncture. For example, processing errors can result from manual editing prior to a machine edit or from the manual correction of computer-generated error flags. The editing process, in general, allows survey managers to review each report for accuracy—an activity that usually results in a feeling of control over the process while obtaining a "sense of the data." While, indeed, there are benefits from editing, recent studies documented by various survey organizations have shown that data are often over-edited (Granquist and Kovar 1997). This over-editing unnecessarily uses valuable resources and can actually add more error to the data than it eliminates.

Processing error can also arise during edit processing due to edit model failure in an automated system. Whether the editing is based on a very sophisticated mathematical model or on simple range checks, the manner in which erroneous data are flagged, and how they are handled, can introduce processing error. For example, if all values for a particular item are forced into an historical range that is no longer reflective of the dynamics of the item, editing error will occur. According to Granquist and Kovar (1997):

> "Over-edited survey data will often lead analysts to rediscover the editor's models, generally with an undue degree of confidence. For example, it was not until a demographer "discovered" that wives are on average two years younger than their husbands that the edit rule which performed this exact imputation was removed from the Canadian Census system!"

Procedures used for editing should be thoroughly documented. This is often difficult and tedious, particularly when the data set contains potentially hundreds of relationships. In addition to documentation about editing procedures, users may be interested in knowing where edits resulted in changes to the reported data, at both a case level and a variable level. Summaries of edits, rates of failure, exceptions, and resolution inconsistencies provide excellent indicators of process quality as do tables of edit changes and percentage distributions of changes by reason for the change. Depending on the number of cases and variables involved, it may be impractical to adequately document this information for every variable. A compromise might be to produce this information for only the most important variables.

Table 7.1, taken from the RECS Quality Profile (U.S. Energy Information Administration 1996), documents the reasons for changes in data cells in the Household File and Billing Files portions of RECS. It provides a useful summary of changes that occurred to the data file and reasons for those changes.

Table 7.1.—1984 Residential Energy Consumption Survey (RECS): Changes to the household and billing files, by reason

	Changes to household file		Changes to billing files	
Reason	Number	Percent	Number	Percent
Total	**20,472**	**100.0**	**4,134**	**100.0**
Keying error	1,868	9.1	([1])	([1])
Coding error	3,699	18.0	([1])	([1])
Clerical error (prior to coding)	([2])	([2])	374	9.0
Interviewer error	1,118	5.5	([2])	([2])
Respondent error	236	1.2	122	3.0
Interviewer or respondent error	422	2.1	([2])	([2])
Data processing error (after keying)	202	1.0	1	([3])
Phone call to respondent household	514	2.5	20	0.5
Phone call to utility/supplier	256	1.3	496	12.0
Other phone call or information	143	0.7	14	0.3
Rental agent (master meter) information	1,251	6.1	—	—
Kerosene survey information	([2])	([2])	—	—
Editor's judgment	9,807	47.8	1,016	24.6
Additional information from questionnaire	545	2.7	25	0.6
None of the above	411	2.0	—	—

— None in this category.

[1] Keying error and coding error were combined (2,066, 50.0 percent) as changes due to keying errors could not be distinguished from changes due to coding errors.

[2] Not applicable.

[3] Less than 0.05 percent.

SOURCE: Residential Energy Consumption Survey Quality Profile (Energy Information Administration 1996).

Even though it does not directly suggest the data are of good quality, thorough documentation is probably a reasonable indicator of a quality data collection system. An example of thorough documentation of a survey's processes, including edits, can be found in the User's Manual for the U.S. Department of Education's 1993–94 Schools and Staffing Survey (Gruber, Rohr, and Fondelier 1996). Another excellent effort of survey documentation was put forth by the U.S. National Center for Health Statistics (1994) when the agency documented the editing for its four families of data systems. Similar efforts are needed for other federal surveys.

One procedure developed to help users assess the extent to which data are altered during the editing stage is to create a shadow variable or flag for each variable (or each critical variable) (Kennickell 1997a and 1997b). The purpose of the shadow variable is to provide information about specific results of the processing procedures. Generally, only a few codes need be used. For example, the shadow variable might contain codes for "as reported," "don't know or refused," "calculated from other variables," "edited to missing," etc. The important feature of the variable is that it should allow users to distinguish item values that are unchanged, edited, or imputed. Once these variables are created, users can employ the shadow variables to determine the amount of “missingness” in a given question and the frequency with which reported values were changed by the processing system. Associated statistics and frequency distributions provide important indicators of data quality.

Shadow variables or flags allow users to determine which cases required the greatest amount of editing. While this approach is still somewhat costly in that the number of variables on the data set doubles, as electronic data storage gets cheaper this becomes less of a factor. Benefits can often far outweigh the costs, as shadow variables can be very useful in tracking changes made to the data. For example, the 1993 National Survey of Small Business Finances, a survey sponsored by the Board of Governors of the Federal Reserve System and the Small Business Administration, associated a shadow variable with each analysis variable. The codes assigned to the shadow variables were values indicating unchanged, reported data; values indicating edited, changed data; and values indicating missing data.

7.2.4 Imputation Errors

While editing and imputation are often thought of as completely distinct activities, in practice they are highly integrated. In some surveys, variables that fail edits are automatically imputed. The error introduced by this practice is part of processing error. However, the same mathematical techniques used to impute variables that are left missing by the respondent (chapter 4) are also used to impute for failed edits. Therefore, imputation error can be discussed as both nonresponse and processing error—depending on the reason the imputed value is needed.

When variables are imputed, the missing or erroneous data are replaced with values generated by an assumed model for the nonrespondent data, so that analyses and summarization can more effectively be performed. For example, in hot deck imputation, the problematic or missing values are replaced by reported values from other units in the survey. Kalton and Kasprzyk (1986) show that many popular imputation methods can be viewed as regression models. If the assumed models do not hold, the imputation process will introduce error. Even if the imputation models are reasonable, imputation often attenuates measures of association among variables.

Documentation of imputation procedures is very valuable for data users. At a minimum, data users should be able to identify the values that have been imputed and the procedure used in the imputation. This permits users the option to employ in their analyses their own methods to compensate for the missing values. The Schools and Staffing Survey (Gruber, Rohr, and Fondelier 1996), for example, developed shadow variables (flags) to indicate both the occurrence of an item imputation and the procedure used (use of administrative data, a hot deck procedure, or a clerical procedure). Similarly, the Federal Reserve Board (Board of Governors of the Federal Reserve System 1997) describes the use of shadow variables, and Kennickell (1998) provides a specific application of their use with imputation. Quality measures available from the use of shadow variables include aggregate statistics on the percentage of missing values and the contribution of imputed values to totals.

Manzari and Della Rocca (1999) proposed an elaborate scheme for measuring the quality of an editing and imputation system for survey data. They do this in the context of evaluating the National Agricultural Statistics Service’s Agricultural Imputation and Edit System (AGGIES). The authors define nine indices of quality. The first three of these assess the quality of the editing, the next three assess the quality of the imputation, and the final three assess the overall quality of both editing and imputation. The indices were computed in the context of a simulation approach in which the authors planted artificial errors in a data set by modifying some of the original values. The modified set of data was processed through the editing/imputation system and the resulting file was compared with the original unmodified file. The indices measured the quality of the editing and imputation procedures based on the number of detected, undetected, and introduced errors. A description of this approach can be found in Todaro and Perritt (2000).

7.3 Reporting Processing Error in Federal Surveys

Data producers usually have quality control systems in place to reduce and monitor data processing errors. To the extent the survey budget allows, useful performance statistics for many of the data processing operations—data entry, coding, and editing—can be developed. These can provide feedback to the survey operations team and ensure that the survey operations meet the data collector's performance objectives. They can even be used to provide performance feedback to individual data entry staff and coders. Probably the most widely implemented performance statistics in federal surveys are those dealing with the editing process. In addition to helping refine edit specifications, performance statistics on editing can help identify data capture problems and even problems with the survey instruments. Edit performance statistics are often relatively inexpensive to implement and can result in substantial survey improvements. This type of information, however, is frequently not reported to users of the data. Users often assume that the quality of data is very high and that strict control systems eliminate processing error. The lack of documentation about processing errors and their consequences may encourage this assumption, when in reality, undocumented processing errors may be important.

Atkinson, Schwanz, and Sieber (1999) found in their review of federal analytic reports that processing was mentioned as a source of error in 78 percent of the publications they reviewed, but that few reports provided any detail. Not one publication presented data keying error rates; only 4 percent presented coding error rates and 6 percent presented edit failure rates.

Specific details of the processing aspects of survey operations are often best described in survey documentation reports (see, for example, Gruber, Rohr, and Fondelier 1996) or special technical reports. For example, the U.S. National Center for Health Statistics (1994) developed a draft report documenting its editing practices across all its survey programs. The Federal Committee on Statistical Methodology (1988) also reported on a major study to evaluate the effect of each processing step on economic census data, by following a set of data items for a sample of establishments through the processing (U.S. Bureau of the Census 1987).

The synthesis of this type of information into a document such as a quality profile has strong appeal for users and provides significant value to the data user community. The quality profiles for the Residential Energy Consumption Survey (U.S. Energy Information Administration 1996) and the American Housing Survey (Chakrabarty and Torres 1996) provide good examples of reports that document processing error. Information contained in such reports documents the quality control aspects of the data processing operations and provides the user with solid information on the efforts of the data collection agency to minimize the various types of error.

The subcommittee recommends that detailed reports such as those cited above be published for ongoing surveys. Analytic reports provide limited information about processing error to the general public. Considerable detail is required to describe the data processing operations and certain performance statistics. While this information is important to survey practitioners and methodologists, it is too detailed to be included in an agency analytic report. Therefore, insofar as the detailed information described above is felt to be beyond the scope of interest for the primary target audience of an analytic report, these statistics and the description of processing operations should be published in a separate volume and referenced by the analytic report. For the analytic report, the subcommittee recommends the following topics for inclusion:

- Errors in data processing ought to be described as a potential source of error in surveys;
- Data keying error rates, scanning error rates, other data entry error rates, coding error rates, and edit failure rates should be referenced as available on the Internet or in technical reports, user manuals, or quality profiles—particularly for key variables;
- A short discussion of the quality control aspects of the data processing operations should be provided to the user; and
- Processing error studies, such as coder-variance studies, should be referenced in the report and be readily available to the user community electronically or through technical reports, user manuals, or quality profiles.

Technical and methodological reports provide the means for disseminating detailed information on all aspects of data processing operations. The subcommittee recommends the following topics be included in more detailed discussions of survey data processing:

- Detailed descriptions of the quality control aspects of survey data processing and the data input operations such as data keying and imaging should be provided;
- Quality control results on data entry, coding, and editing should be reported;
- Processing error studies, particularly coder-variance studies, should be described, the results summarized, and implications, if any, for analysis clarified;
- The extent to which data processing operations alter responses, particularly in the edit and imputation phases, should be discussed. The altered data should be identified for the end-users in any data files provided; and
- In continuing and periodic surveys, changes in processing operations should be identified and described.

References

Appel, M.V. and Hellerman, E. 1983. "Census Bureau Experience with Automated Industry and Occupation Coding." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 32–40.

Atkinson, D., Schwanz, D., and Sieber, W.K. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Biemer, P. and Caspar, R. 1994. "Continuous Quality Improvement for Survey Operations: Some General Principles and Applications." *Journal of Official Statistics*. 10: 307–326.

Blom, E. and Lyberg, L. 1998. "Scanning and Optical Character Recognition in Survey Organizations." In M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls, and J.M. O'Reilly (eds.), *Computer Assisted Survey Information Collection*. New York: John Wiley & Sons. 499–520.

Board of Governors of the Federal Reserve System. September 19, 1997. *Codebook for the 1993 National Survey of Small Business Finances.* Mimeo. 3, 11, 17, and 27. Available at http://www.bog.frb.fed.us/boarddocs/surveys

Chakrabarty, R.P. and Torres, G. 1996. *American Housing Survey: A Quality Profile.* Washington, DC: U.S. Department of Housing and Urban Development and U.S. Department of Commerce (Current Housing Reports, H121/95-1).

Dielman, L. and Couper, M. 1995. "Data Quality in a CAPI Survey: Keying Errors." *Journal of Official Statistics.* 11: 141–146.

Federal Committee on Statistical Methodology. 1990. *Data Editing in Federal Statistical Agencies.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 18).

Federal Committee on Statistical Methodology. 1988. *Quality in Establishment Surveys.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 15).

Granquist, L. and Kovar, J. 1997. "Editing of Survey Data: How Much is Enough?" In L. Lyberg, P. Biemer, M. Collins, E.D. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality.* New York: John Wiley & Sons. 415–435.

Gruber, K., Rohr, C., and Fondelier, S. 1996. *1993–94 Schools and Staffing Survey Data File User's Manual, Volume 1: Survey Documentation.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 96–142).

Henderson, L. and Allen, D. 1981. *NLS Data Entry Quality Control: The Fourth Follow-up Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Jabine, T. J. December 1987. *Review of Computer Edit and Update Performance Statistics for the Residential Energy Consumption Survey, Final Report.* Washington, DC.

Jabine, T., King, K., and Petroni, R. 1990. *SIPP Quality Profile*. Washington, DC: U.S. Bureau of the Census.

Jones, M. and Green, I. 1998. "Utilization of Document Imaging Technology by the 1996 Canadian Census of Agriculture." *Conference Proceedings of Agricultural Statistics 2000.* Washington, DC.

Kalton, G. and Kaspryzk, D. 1986. "The Treatment of Missing Survey Data." *Survey Methodology.* 12: 1–16.

Kalton, G. and Stowell, R. 1979. "A Study of Coder Variability." *Journal of the Royal Statistical Society*, Series C (Applied Statistics). 28(3): 276–289.

Kennickell, A.B. September 1998. *Multiple Imputation in the Survey of Consumer Finances.* Board of Governors of the Federal Reserve System (working paper).

Kennickell, A.B. June 3, 1997a. *Codebook for the 1995 Survey of Consumer Finances.* Mimeo, Board of Governors of the Federal Reserve System.

Kennickell, A.B. January 1997b. *Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances.* Mimeo, Board of Governors of the Federal Reserve System.

Linacre, S.J. 1991. "Approaches to Quality Assurance in ABS Business Surveys." *Proceedings of the International Statistical Institute (ISI), 48th Session.* 2: 487–511.

Linacre, S.J. and Trewin, D.J. 1989. "Evaluation of Errors and Appropriate Resource Allocation in Economic Collections." *Proceedings of the Annual Research Conference.* Washington, DC: U.S. Bureau of the Census. 197–209.

Lyberg, L. and Kasprzyk, D. 1997. "Some Aspects of Post-Survey Processing." In L. Lyberg, P. Biemer, M.Collins, E.D. deLeeuw, C. Dippo, N. Schwarz, and D.Trewin (eds.), *Survey Measurement and Process Quality.* New York: John Wiley & Sons. 353–370.

Manzari, A. and Della Rocca, G. 1999. *A Generalized System Based on a Simulation Approach to Test the Quality of Editing and Imputation Procedures*. Conference of European Statisticians, Work Session on Statistical Data Editing, United Nations Statistical Commission and Economic Commission for Europe (Working Paper No. 13).

Morganstein, D. and Marker, D. 1997. "Continuous Quality Improvement in Statistical Agencies." In L. Lyberg, P. Biemer, M. Collins, E.D. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons. 475–500.

Todaro, T. and Perritt, K. 2000. "Overview and Evaluation of AGGIES, an Automated Edit and Imputation System." *1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 30). 481–489.

U.S. Bureau of the Census. 1987. *1982 Economic Censuses and Census of Governments Evaluation Studies*. Washington, DC.

U. S. Bureau of the Census. October 18, 1993. “Memorandum for Thomas C. Walsh from John H. Thompson, Subject: 1990 Decennial Census-Long Form (Sample Write-In) Keying Quality Assurance Evaluation.” M. Roberts author.

U.S. Bureau of the Census. 1998. *Survey of Income and Program Participation (SIPP) Quality Profile*. 3rd Edition. Washington, DC.

U.S. Energy Information Administration. 1996. *Residential Energy Consumption Survey Quality Profile*. Washington, DC.

U.S. National Center for Health Statistics. 1994. *Data Editing at the National Center for Health Statistics*. Internal draft report available from Kenneth Harris, chair of the NCHS Editing Committee.

Chapter 8

Total Survey Error

8.1 Introduction

In the previous chapters, a variety of systematic and variable errors that could distort the distribution of survey estimates were discussed. A natural end to the investigation of the important sources of error in a survey is the integration of these separate errors into an estimate of the overall or *total survey error* in estimates computed from the survey.

Estimates of total survey error are clearly of value for data users. Commonly, users rely on estimated standard errors, reflecting only errors due to sampling, to make statistical inferences such as confidence intervals and tests of hypotheses. An estimate of total survey error that accounts for sampling and nonsampling errors, both systematic and variable, would be more appropriate for use in these situations.

Survey designers would find total survey error estimates of immense value in improving survey methods. Estimates of total survey error could pinpoint areas of the survey that most need improvement and efforts could be concentrated on those areas. Thus, total survey error estimates could help determine how much effort should be placed on improving different aspects of the survey process, such as sample design, questionnaires, interviewer training, coding, and processing.

The concept of total survey error seems obvious—total survey error is the cumulative effect that all sources of error in a survey have on the distribution of the estimates. Operationalizing this idea in a manner to produce estimates of this error is nevertheless difficult. Total survey error often is formulated by survey statisticians in terms of the mean squared error of the estimate, where the mean squared error is the sum of the variance and the square of the bias. However, this formulation does not capture the complexity of the problem. For example, consider the interviewer as a source of error in addition to sampling error. Interviewer errors can result in both bias (systematic error) and variance (variable error). Since the same source of error contributes to both terms, the simple additive structure of the mean squared error may not be adequate as a model for estimating total survey error.

Estimating total survey error is difficult because it is intrinsically a multivariate problem, where each error source contributes to the systematic and variable error of the estimate. The important sources of error also vary from survey to survey. For example, coverage and interviewer errors might be the most critical error sources in one survey, while in another survey the most important sources of errors might be questionnaire context effects, time in-sample, nonresponse, or recall errors. The method of estimating total survey error must be flexible enough to allow different errors to be incorporated.

The multivariate nature of estimating total survey error causes greater difficulty because the component error sources may be correlated. For example, in surveys that ask about income, two important error sources might be nonresponse and response errors. Nonrespondents are often concentrated at both extremes of the income distribution. There is also considerable evidence that the response errors for wealthy and poor persons who do respond to the survey are very different. In this situation it is easy to postulate a correlation between the nonresponse error and response error in the estimation of income. As the dimensions of the problem grow with more error sources, the difficulty in estimating the covariance structure increases. Assuming uncorrelated error sources makes the problem much more tractable, but such simplifying assumptions are not consistent with the results of empirical research.

One other aspect of the multivariate nature of the problem of estimating total survey error is that most surveys produce many estimates and these estimates may be affected differently by the error sources. For example, the effect of coverage error in estimating a characteristic may be very large, but when the estimated characteristics for two subgroups are compared this error may be less important than others. Thus, the estimates of total survey error should be able to account for these dimensions of the problem.

The discussion of estimating total survey error thus far has assumed implicitly that a mathematical model exists that includes all the key error sources of the survey. Pioneering efforts that developed more inclusive survey error models have done exactly this. Some early examples of this approach are Hansen, Hurwitz, and Bershad (1961); Kish (1965); Fellegi (1964); and Sukhatme and Seth (1952). The use of models and their limitations in the estimation of total survey error are discussed in more detail later.

Another way of approaching total survey error is to examine the key processes in a particular survey and describe what is known or suspected about the potential error contributed by these sources. This is what quality profiles do; they inform about all potential error sources in the survey. The first quality profile by Brooks and Bailar (1978) was called an error profile, highlighting the focus on all sources of error in the survey. Brooks and Bailar examined a single statistic from one survey—employment estimated from the Current Population Survey (CPS).

The goal of quality profiles is to enhance data users' understanding of the limitations of the statistics produced from the survey and to guide producers in improving survey operations, and in turn, improve the quality of the statistics. The content of a typical quality profile covers topics such as coverage, nonresponse, measurement error, data processing, and estimation, and includes quantitative as well as qualitative results. A review of quality profiles and their role in providing information about total survey error is given later.

Another commonly used method of evaluating the quality of survey data is to compare survey estimates to statistics from independent sources. This approach is very different from that of quality profiles which examine and explore the survey processes for potential sources of error. Instead, the comparison of estimates to independent sources focuses on the aggregate error irrespective of the source.

The goal of comparing estimates from surveys to independent sources is the same as that of quality profiles, but the emphasis is different. In quality profiles, the processes are closely examined and tend to provide greater information on which processes need to be improved. On the other hand, comparisons to independent sources concentrate on how well the statistics match up and may be of primary benefit to data users. Of course, both approaches do provide insight for both of these goals and the difference is a matter of emphasis.

The most challenging aspect of making comparisons to independent sources is developing comparable data and interpreting the differences as a measure of total survey error. Whether the independent source is administrative records or another survey, the estimates are often not comparable for a variety of reasons. Even when comparable data can be found, it is still difficult to interpret differences between the survey and the independent source as a measure of total survey error in the survey estimate because the independent source is never error free. This is discussed further below.

No completely satisfactory method exists for estimating total survey error. However, this should not discourage attempts to measure it, even if the measures are imperfect. Virtually every effort to take a more rounded and complete view of survey error has resulted in improvements in both survey methods and understanding of the limitations of the data by users. The practice of integrating what is known about errors in a survey in some manner, whether it is a quantitative estimate of total survey error or a qualitative evaluation of potential errors, is valuable.

8.2 Measuring Total Survey Error

8.2.1 Comparisons to Independent Sources

Comparing estimates from a survey to values from independent data sources is a useful method of examining the overall effect of errors on the estimates from the survey, but it is difficult to quantify the benefits. In most cases, the comparisons give a broad overview of the cumulative effect of errors in the survey. In a few cases, comparisons may reveal areas that need to be investigated further and this may lead directly to improvements in the survey procedures or methods.

One of the primary beneficiaries of comparisons to independent sources are data users, especially those who are familiar with statistics produced from the independent sources. The comparisons provide users with insights into how the statistics from the survey align to statistics from other sources and highlight potential differences that might otherwise cause confusion.

The analysis and reporting of comparisons to independent sources is an important ingredient in assessing total survey error. However, these comparisons are not always released to the public, as is reported by the Federal Committee on Statistical Methodology (1988). When the comparisons are released, data producers have done so in a variety of formats. Kim et al. (1996) and Nolin et al. (1997) use a working paper format to compare a variety of statistics from the 1995 and 1996 National Household Education Survey to data from multiple independent sources. Vaughan (1988 and 1993) provides detailed aggregate comparisons of income statistics from the 1984 CPS, Survey of Income and Program Participation (SIPP), and administrative program data in a conference proceedings paper and in a technical report. The Energy Information Administration (EIA) publishes data comparisons as feature articles in its monthly publications (U.S. Energy Information Administration 1999a). Other formats for the release of the comparisons include chapters in quality profiles and appendices in survey reports.

The independent sources of statistics may be either administrative records prepared for nonstatistical purposes or estimates from other surveys. Administrative data or data from program sources are often viewed as more accurate than survey estimates; thus, comparisons to these sources may be used to measure the total survey error. This supposition, however, is not always true because administrative records are frequently not edited or subjected to other assessments as discussed in the previous chapters. Similarly, when the independent source is another survey, care must be taken in evaluating the differences because both surveys may have different error structures.

A key aspect of the evaluation of total survey error by comparing the survey estimates to independent sources is the issue of comparability. If the statistic from the independent data source is error-free, then the difference between it and the survey statistic accurately estimates total survey error. In practice, error-free statistics from independent sources are virtually nonexistent. Nonetheless, the comparisons are still useful when the independent data source has a low level of error compared to the error in the survey estimate. In this case, the difference between the survey statistic and the independent source may be a useful indicator of the direction and magnitude of the total survey error.

When comparisons are made to independent data sources, the error sources of the independent data must be taken into account along with the error sources of the survey. The most common factors that must be considered when comparisons are made include: the time period of the data collection, coverage errors, sampling errors, nonresponse errors, processing errors, measurement errors, and mode effects. Many of the effects of these factors are easy to understand and need little or no explanation. For example, comparisons between estimates of the number of persons in the United States who were not born in the United States from the 1993 CPS and the 1990 Decennial Census will differ due to changes between the two time periods alone. The way some of these factors might affect comparisons are less obvious and are discussed below.

Coverage is an important aspect of comparability. For example, in random digit dial (RDD) telephone surveys only telephone households are covered so comparisons between RDD survey estimates and independent data sources may be informative about the nature and size of coverage biases in the RDD survey. Nolin et al. (1997) compared estimates from a national RDD survey, the National Household Education Survey, and an independent source from a survey conducted in both telephone and nontelephone households, the CPS. In the review, Nolin et al. (1997) noted that about 6 percent of adults aged 16 years and older who were not enrolled in elementary or secondary school lived in nontelephone households while about 10 percent of the children under 11 years old lived in nontelephone households. The comparisons were used to examine whether the estimates from the RDD survey which were statistically adjusted to reduce coverage biases differ significantly from the CPS estimates. Since the CPS estimates did not have this particular coverage bias, the absence of significant differences was taken as evidence of low levels of coverage bias in the estimates from the RDD survey.

If comparisons with external sources are made regularly over time, a change in the difference between two series may signal a change in the industry that needs to be resolved (coverage error). The EIA regularly compares its petroleum supply data to any available related information (U.S. Energy Information Administration 1999b). In the 1980's, the EIA compared fuel data with data from administrative records at the U.S. Department of Transportation, and in the regular comparison a 4 percent bias had always existed. When the bias increased to 7 percent, research showed that the industry had changed and new types of companies were producing gasoline by methods not covered by the EIA's survey.

The need to consider errors in both the survey and the independent data source is clearly demonstrated when the independent source is subject to sampling errors. In this situation, differences between the survey estimates and the independent data source have variances that are the sum of the variance due to the two sources. For example, Coder and Scoon-Rogers (1996) give differences between March CPS and SIPP estimates of income. Both surveys are affected by sampling error, so any differences in the estimates from these surveys are also affected by this error.

Other chapters in this volume have discussed different sources of measurement error in surveys and these apply equally to the survey and the independent source. Each potential error source, the interview, the questionnaire, the data collection method, and the respondent, ought to be considered. For example, administrative program data may collect family income to determine eligibility for a federal assistance program, but the program may define the concept of family differently from the survey. This could lead to differences that are not due to error in the survey, but to differences in the definitions. Another example is that program data and surveys may count different units: the survey may count persons and the program data may count households.

In summary, comparing survey estimates to independent sources can provide valuable information about the nature of the total survey error in the estimates, but the benefits are limited because the independent sources are also subject to error. These comparisons are most valuable when the statistics from the independent source are highly accurate. When this is the case, the differences observed can be considered valid measures of the total survey error. When the independent source has significant error of its own, the differences may still reveal important features of the survey that data users would find useful even though the differences cannot be considered valid measures of total survey error.

8.2.2 Quality Profiles

For the past 20 years, the main approach to providing information about total survey error has been the development and publication of survey-specific quality profiles. The origins and evolution of quality profiles are described by Jabine (1991). Three main functions of quality profiles are to provide qualitative and quantitative information about total survey error and the principal components of those errors; to summarize research and information on the quality of a survey; and to give a systematic account of error sources that affect estimates, which can then be used to direct improvement activities.

Zarkovich (1966) was an early attempt to accomplish many of these objectives, but Brooks and Bailar (1978) is generally recognized as the prototype of the modern quality profile. Even Brooks and Bailar (1978) differ from most of the later quality profiles because they concentrated on one statistic while most current quality profiles are concerned with all the statistics produced from the survey. The value of quality profiles is evident from the proliferation of quality profiles in recent years. Quality profiles on federal surveys include ones done for the Survey of Income and Program Participation (U.S. Bureau of the Census 1998), the American Housing Survey (Chakrabarty and Torres 1996), the Residential Energy Consumption Survey (U.S. Energy Information Administration 1996), and the Schools and Staffing Survey (Kalton et al. 2000). In addition to these, other closely related efforts include assessments of the quality of establishment surveys (Federal Committee on Statistical Methodology 1988), an error profile on multiple frame surveys (Beller 1979), and a study of selected nonsampling errors in a survey of recent college graduates (Brick et al. 1994).

Most quality profiles are developed for periodic and continuing surveys and written to support the needs of both data users and those responsible for the operations of the survey. Data users find quality profiles valuable because they are summaries of research about the quality of the survey and this helps users better understand issues that should be addressed when the data are analyzed. Data producers value quality profiles because they provide succinct and wide-ranging documentation. It should be noted, though, that quality profiles are distinctly different from survey documentation that is usually presented in the form of a User's Manual. Quality profiles are used to direct new research to improve the quality of the data, and secondarily to initiate new staff into the operations, design and principal error sources of the survey. Agencies developing and implementing new surveys with similar characteristics also find quality profiles useful at the design stage. For example, the SIPP quality profile is a valuable document for those developing longitudinal household surveys.

Quality profiles generally cover a large number of potential sources of error and thereby deal with total survey error rather than focusing on simply one type of survey error. The structure of most quality profiles is very similar, with sections devoted to sample design, data collection, and estimation. Often whole chapters are devoted to sources suspected of having the largest potential effect on the statistics produced from the survey. For example, the RECS quality profile (U.S. Energy Information Administration 1996) has separate chapters devoted to coverage, nonresponse, and measurement error, three of the largest potential sources of error in the survey. In common with most other quality profiles it also has an overview of the survey, a discussion of data processing, imputation, estimation and sampling error, and a comparison to independent data sources.

Most of the quality profiles contain many tables and charts that quantify specific aspects of the components of total survey error. References to the original documents with more complete reports on the specific errors are also given. These summaries and references are valuable and help identify the error sources that warrant research. In a study of selected nonsampling errors in a survey of recent college graduates, Brick et al. (1994) attempt to synthesize the errors from different sources and make more general statements about the total survey error. This represents a departure from the content of most quality profiles in that most profiles compile information about a survey's quality components without attempting to tie these components together to make a statement about the total effects. In the next section, the framework for taking this additional step is outlined along with some of the limitations of doing so.

8.2.3 Error Models

When survey estimates are compared to independent data sources, the comparisons are nearly always at an aggregated level and provide little or no information on the specific source of the differences at the unit level. If the differences are small, then data users may be more comfortable using the survey data for estimation of these and other quantities. However, comparisons with independent sources do not permit direct evaluations of sources of errors in the survey that are needed to understand and improve the survey process. Furthermore, these types of comparisons usually focus on biases and rarely even consider variable errors.

Quality Profiles suffer from the opposite problem—virtually every quality profile is analytic in the sense that it describes each key process in the survey in isolation from the other processes in the survey. Another concern with quality profiles is that they frequently do not provide quantitative estimates of the errors in estimates due to different sources. For many sources of error, only descriptive accounts are offered. Even when specific estimates of errors from different sources are given, quality profiles do not suggest methods for understanding how the disparate errors affect the overall distribution of errors in the estimates.

Estimation of total survey error requires a method of synthesizing what is known about errors arising from different error sources. Researchers attempting this synthesis have proposed mathematical models that incorporate a variety of sources of error at the individual unit level.

Early in the development of this approach Hansen, Hurwitz, and Bershad (1961) suggested a model that included the interviewer as a source of error in addition to simple response bias and variance. This model has been very influential and has served as the basis for many extensions. The model also had great practical importance because estimates from the model revealed that interviewers were an important source of the total error in estimates computed from the decennial censuses. Consequently, the census was changed to self-administered mail interviews.

Despite the usefulness of this model, it does not provide an estimate of total survey error. It includes only the respondent and the interviewer as sources of error, failing to capture sources such as nonresponse, coverage, questionnaire wording, context, and other field and processing sources of error.

A related model was proposed by Kish (1965) to incorporate various sources of bias and variable error in estimates from a survey. This model was used by Andersen et al. (1979) in their exploration of errors in a survey of health services use and expenditures. In this case, the model was focused on three components of nonsampling error: nonresponse bias, field bias (typically denoted as response bias today), and processing bias (limited to imputation bias). The biases were estimated by comparing the survey responses to external data sources and assuming differences from the external data sources were biases.

Even though this effort provided important quantitative insights into key error sources for this survey, the approach fails to cover important error sources in many surveys. It also relies on external sources of data that can be assumed to contain the true characteristic of the unit; such resources are not available for most items in surveys. Another shortcoming of the model is that it does not deal with measuring variable errors, except as measured by the traditional sampling error.

Bailar and Biemer (1984) extended Hansen, Hurwitz, and Bershad's model to explicitly address nonsampling error. The extended model is more complete, but it is difficult to estimate the parameters of the model because of interactions in the errors. It is possible to estimate the parameters if the interaction terms are dropped from the model, but this assumption is not supported by many other research findings. As a result, Bailar and Biemer (1984) did not actually compute any estimates from this more complete model of survey error.

Several other examples of mathematical models that incorporate different sources of error have been presented.

- Groves and Magilavy's model (1984) includes sources of error for sampling, refusing to be interviewed, other reasons for noninterviews, and response errors. They simplify the model by assuming the interactions are all zero and use record check data to compute estimates of error for 20 statistics.

- Woltman and Johnson (1989) give an extensive mathematical model of survey error for the 1990 census, but do not apply the model to produce estimates of total survey error.

- Mulry and Spencer (1990) examine the use of Bayesian models to estimate the error in the dual-system estimate of the total population based on post-enumeration survey data. A consequence of using Bayesian methods is that Mulry and Spencer (1990) can synthesize the different errors from the posterior distribution of the net undercount rate.

- Alwin (1991) changes the perspective on total survey error, by starting with a population model of measurement error and layers other errors, such as nonresponse and sampling error, on top of this structure.

Brick et al. (1994) studied important sources of error in estimating characteristics from the National Center for Education Statistics 1991 Survey of Recent College Graduates (RCG). In order to estimate total survey error, complex models such as those described by Bailar and Biemer (1984) were investigated, but the problem of estimating the model parameters without oversimplifying assumptions could not be resolved. Because of this difficulty, a less structured approach was attempted.

Rather than specify a given model, the joint effect of the errors from each of the investigated sources of error was assessed for particular examples. In each example, evidence from the analysis of each error source studied was considered and correlations between the sources were conjectured. Based on this informal approach, recommendations were presented to help users estimate the magnitude of error in the estimates. The examples included overall estimates of the percentage of graduates with a characteristic (the percentage certified to teach) and comparisons between subgroups (the difference between the percentage of males and females working for pay).

Brick et al. (1994) recommended that it was not advisable to make any adjustments in the estimates for nonresponse bias, because these biases could not be estimated well from the data available. Similarly, response bias was estimated from a relatively small subsample (500 of 12,000 respondents were included in the study) and adjusting the estimates from such a small subsample would lead to very large variable errors.

An adjustment that was recommended was to use more conservative statistical inference procedures, such as using a 99-percent confidence interval in place of 95-percent intervals, for estimates that are most affected by survey errors. For example, the correlated response variance due to interviewers might result in substantially larger variances than the sampling variances computed from the survey indicate. For such statistics, using 99-percent confidence intervals would provide more protection against making erroneous Type I errors. For other estimates that are not as greatly affected by the errors, conservative statistical inference procedures might not be needed.

8.3 Reporting Total Survey Error

The study of total survey error is an important effort that can result in improvements in the conduct and analysis of federal government surveys. Understanding the individual sources of error in the surveys and their contribution to the overall error are key ingredients in improving the way surveys are designed and implemented. Knowledge of the effects of error sources on statistics computed from the survey can lead to the use of analytic methods that are robust against such errors.

The three main approaches to studying total survey error in federal government surveys are comparison to independent data sources, development and dissemination of quality profiles, and estimation of error components using models of the errors. Each of the methods has advantages and disadvantages outlined earlier in this chapter.

The assessment of total survey error is not a simple task. Miller (1997) discusses some of the difficulties encountered in trying to create a summary measure of quality for surveys of the EIA. Miller describes the evolution of approaches used and concludes that, despite the best of efforts, reasonable summary measures of total survey error pose formidable obstacles in their development and computation.

Bailar (1983) also examines ways of assessing total survey error, concentrating primarily on the slowness of the production of quality profiles for federal government surveys. Nearly 20 years later, her observations remain relevant. While the number of surveys with quality profiles has increased since the early 1980s, surveys with quality profiles remain the exception rather than the rule.

As is clear from the works cited in this chapter, substantial technical and practical problems exist in the measurement of total survey error. From a survey program's point of view, the complexity and time-consuming nature of the work on total survey error usually is weighed against the more immediate perceived needs of the survey programs. In the subcommittee's review of agency reporting practices, only a limited number of studies have been identified. This is unfortunate because, as Bailar (1983) points out, efforts aimed at understanding survey error often uncover sources of errors that can be easily remedied yielding improved statistics. She also suggests that efforts to better understand total survey error are a critical stimulus to improving quality. Steps taken to study error sources and their relationships to each other and the statistics from a survey nearly always increase the chances of improving the quality in that survey.

Measurement issues result in obvious difficulties in the reporting of total survey error. The sources of information on total survey error are refereed journals, quality profiles, research working papers, and conference presentations. The subcommittee found a limited number of such studies. In the course of the subcommittee's review, however, several other reporting mechanisms, usually less comprehensive than a report on total survey error, were identified that are important contributions to the reporting of information on sources of error in surveys. For example, the technical paper describing the design and methodology of the Current Population Survey (U.S. Bureau of the Census and U.S. Bureau of Labor Statistics 2000) provides detailed documentation of the survey design, estimation, and data collection procedures. In addition, however, the report provides chapters on data quality concepts, sources and control of nonsampling errors, and quality indicators of nonsampling error. While identifying itself as a technical report on the design and methodology of the Current Population Survey, the report provides an enormous amount of information in the spirit of a quality profile.

The U.S. National Center for Health Statistics (1994) has drafted detailed documentation of the editing systems used in its survey programs. Similarly, the NCES (Salvucci et al. 1997) reviews measurement error studies conducted by its survey programs. Thus, a specific source of error is reviewed across the agency's survey programs. So while total survey error is not necessarily frequently reported, specific error sources are, in fact, studied by some agencies. The subcommittee's recommendations, with respect to reporting total survey error, reflect some of the good practices observed in federal statistical agencies.

The recommendations are:

- Continuing and periodic survey programs ought to regularly report a summary of the results of methodological studies related to a survey program. The implications of these results on analysis should be addressed. One way to address this point is through the development of quality profiles for continuous and periodic survey programs.

- Lacking a report on all error sources, survey programs should identify their most prominent sources of error, report on them individually in technical reports, and discuss the implications for analysis.

- Production or process statistics should be routinely calculated, discussed, and reported to survey data users**.**

- Survey programs should allocate a portion of their budgets to design, implement, and report on methodological studies that assist the user to understand the sources of error in the survey and their implications for analysis.

- Technical or methodology reports that describe in some detail the sample design, estimation, and data collection procedures should be developed for data users. In the absence of specific methodological studies, such reports help users to make judgements concerning the quality of the survey and its operations.

- Survey programs should compare their aggregate results to other comparable data, assess reasons for differences, and report on the results of any comparisons.

- Web applications should be developed to allow the data user to easily navigate through information concerning the overall quality of the data. Continuous and periodic surveys should develop long-range research plans that systematically address the measurement of the components of total survey error.

References

Alwin, D. 1991. "Research on Survey Quality." *Sociological Methods and Research.* 20: 3–29.

Andersen, R., Kasper, J., Frankel, M., and Associates. 1979. *Total Survey Error—Applications to Improve Health Surveys.* San Francisco, CA: Jossey-Bass.

Bailar, B. 1983. "Error Profiles: Uses and Abuses." In T. Wright (ed.), *Statistical Methods and the Improvement of Data Quality.* New York: Academic Press. 117–130.

Bailar, B. and Biemer, P. 1984. "Some Methods for Evaluating Nonsampling Error in Household Censuses and Surveys." In P.S.R.S Rao and J. Sedransk (eds.), *W.G. Cochran's Impact on Statistics.* New York: John Wiley & Sons. 253–274.

Beller, N. 1979. *Error Profile for Multiple Frame Surveys.* Washington, DC: U.S. Department of Agriculture, Statistical Reporting Service Research Report.

Brick, J.M., Cahalan, M., Gray, L., and Severynse, J. 1994. *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (Technical Report NCES 95–640).

Brooks, C.A. and Bailar, B.A. 1978. *An Error Profile: Employment as Measured by the Current Population Survey.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 3).

Chakrabarty, R.P. and Torres, G. 1996. *American Housing Survey: A Quality Profile*. Washington, DC: U.S. Department of Housing and Urban Development and U.S. Department of Commerce. (Current Housing Reports H121/95-1).

Coder, J. and Scoon-Rogers, L. 1996. *Evaluating the Quality of Income Data Collected in the Annual Supplement to the March Current Population Survey and the Survey of Income and Program Participation.* Washington, DC: U.S. Bureau of the Census (Working Paper 96–04).

Federal Committee on Statistical Methodology. 1988. *Quality in Establishment Surveys.* Washington, DC: U.S. Office of Management and Budget (Statistical Working Paper 15).

Fellegi, I. 1964. "Response Variance and its Estimation." *Journal of the American Statistical Association*. 59: 1,016–1,041.

Groves, R. and Magilavy, L. 1984. "An Experimental Measurement of Total Survey Error." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 698–703.

Hansen, M.H., Hurwitz, W.N., and Bershad, A. 1961. "Measurement Errors in Censuses and Surveys." *Bulletin of the International Statistical Institute.* 38: 359–374.

Jabine, T. 1991. "The SIPP Quality Profile." *Seminar on the Quality of Federal Data, Part 1 of 3*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 20). 19–28.

Kalton, G., Winglee, M., Krawchuk, S., and Levine, D. 2000. *Quality Profile for SASS: Rounds 1–3: 1987–1995*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 2000–308).

Kim, K., Loomis, L., Collins, M., and Chandler, K. 1996. *Comparison of Estimates from the 1995 National Household Education Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (Working Paper 96–30).

Kish, L. 1965. *Survey Sampling.* New York: John Wiley & Sons.

Miller, R. 1997. "Data Quality at the Energy Information Administration: The Quest for a Summary Measure." *Seminar on Statistical Methodology in the Public Service.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 26). 145–156.

Mulry, M. and Spencer, B. 1990. "Total Error in Post Enumeration Survey (PES) Estimates of Population: The Dress Rehearsal Census of 1988." *Proceedings of the Census Bureau Annual Research Conference.* Washington, DC: U.S. Bureau of the Census. 326–359.

Nolin, M.J., Collins, M.A., Vaden-Kiernan, N., Davies, E., and Chandler, K. 1997. *Comparison of Estimates in the 1996 National Household Education Survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics (Working Paper 97–28).

Salvucci, S., Walter, E., Conley, V., Fink, S., and Saba, M. 1997. *Measurement Error Studies at the National Center for Education Statistics*. Washington, DC: U.S. Department of Education, National Center for Education Statistics (NCES 97–464).

Sukhatme, P.V. and Seth, G.R. 1952. "Nonsampling Errors In Surveys." *Journal of the Indian Society of Agricultural Statistics.* 4: 5–41.

U.S. Bureau of the Census. 1998. *Survey of Income and Program Participation (SIPP) Quality Profile*. 3rd Edition. Washington, DC: U.S. Department of Commerce.

U.S. Bureau of the Census and U.S. Bureau of Labor Statistics. 2000. *Current Population Survey: Design and Methodology*. Washington, DC: U.S. Department of Commerce (Technical Paper 63).

U.S. Energy Information Administration. 1999a. "Comparisons of Independent Petroleum Supply Statistics." *Petroleum Supply Monthly*. Washington, DC: U.S. Department of Energy (DOE/EIA-0109 [1999/12]).

U.S. Energy Information Administration. 1999b. "A Comparison of Selected EIA-782 Data with Other Data Sources." *Petroleum Marketing Monthly*. Washington, DC: U.S. Department of Energy (DOE/EIA-0380 [1999/12]).

U.S. Energy Information Administration. 1996. *Residential Energy Consumption Survey Quality Profile.* Washington, DC.

U.S. National Center for Health Statistics. 1994. *Data Editing at the National Center for Health Statistics*. Internal draft available from Kenneth Harris, Chair of the NCHS Editing Committee. Hyattsville, MD.

Vaughan, D. 1988. “Reflections on the Income Estimates from the Initial Panel of the Survey of Income and Program Participation.” *Individuals and Families in Transition: Understanding Change Through Longitudinal Data.* Washington DC: U.S. Bureau of the Census. 333–416.

Vaughan, D. 1993. *Reflections on the Income Estimates from the Initial Panel of the Survey of Income and Program Participation (SIPP), Studies in Income Distribution.* Washington, DC: U.S. Department of Health and Human Services, Social Security Administration (No. 17, SSA Pub. No. 12-11776 (17)).

Woltman, H. and Johnson, R. 1989. “A Total Error Model for the 1990 Census.” *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 522–527.

Zarkovich, S. 1966. *Quality of Statistical Data.* Food and Agricultural Organization of the United Nations: Rome, Italy.

Reports Available in the Statistical Policy Working Paper Series

1. *Report on Statistics for Allocation of Funds*, 1978 (NTIS PB86-211521/AS)
2. *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, 1978 (NTIS PB86-211539/AS)
3. *An Error Profile: Employment as Measured by the Current Population Survey*, 1978 (NTIS PB86-214269/AS)
4. *Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics*, 1978 (NTIS PB86-211547/AS)
5. *Report on Exact and Statistical Matching Techniques*, 1980 (NTIS PB86-215829/AS)
6. *Report on Statistical Uses of Administrative Records*, 1980 (NTIS PB86-214285/AS)
7. *An Interagency Review of Time-Series Revision Policies*, 1982 (NTIS PB86-232451/AS)
8. *Statistical Interagency Agreements*, 1982 (NTIS PB86-230570/AS)
9. *Contracting for Surveys*, 1983 (NTIS PB83-233148)
10. *Approaches to Developing Questionnaires*, 1983 (NTIS PB84-105055)
11. *A Review of Industry Coding Systems*, 1984 (NTIS PB84-135276)
12. *The Role of Telephone Data Collection in Federal Statistics*, 1984 (NTIS PB85-105971)
13. *Federal Longitudinal Surveys*, 1986 (NTIS PB86-139730)
14. *Workshop on Statistical Uses of Microcomputers in Federal Agencies*, 1987 (NTIS PB87-166393)
15. *Quality in Establishment Surveys*, 1988 (NTIS PB88-232921)
16. *A Comparative Study of Reporting Units in Selected Employer Data Systems*, 1990 (NTIS PB90-205238)
17. *Survey Coverage*, 1990 (NTIS PB90-205246)
18. *Data Editing in Federal Statistical Agencies*, 1990 (NTIS PB90-205253)
19. *Computer Assisted Survey Information Collection*, 1990 (NTIS PB90-205261)
20. *Seminar on Quality of Federal Data*, 1991 (NTIS PB91-142414)
21. *Indirect Estimators in Federal Programs*, 1993 (NTIS PB93-209294)
22. *Report on Statistical Disclosure Limitation Methodology*, 1994 (NTIS PB94-165305)
23. *Seminar on New Directions in Statistical Methodology*, 1995 (NTIS PB95-182978)
24. *Electronic Dissemination of Statistical Data*, 1995 (NTIS PB96-121629)
25. *Data Editing Workshop and Exposition*, 1996 (NTIS PB97-104624)
26. *Seminar on Statistical Methodology in the Public Service*, 1997 (NTIS PB97-162580)
27. *Training for the Future: Addressing Tomorrows Survey Tasks*, 1998 (NTIS PB99-102576)
28. *Seminar on Interagency Coordination and Cooperation*, 1999 (NTIS PB99-132029)
29. *Federal Committee on Statistical Methodology Research Conference (Conference Papers*), 1999 (NTIS PB99-166795)
30. *1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings*, 2000 (NTIS PB2000-105886)
31. *Measuring and Reporting Sources of Error in Surveys*, 2001 (NTIS PB2001-104329)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; telephone: 1–800–553–6847. The Statistical Policy Working Paper series is also available electronically from FCSM's web site < http://www.fcsm.gov>.