Margins of Error

A Study of Reliability in Survey Measurement

Duane F. Alwin
McCourtney Professor of Sociology and Demography
Pennsylvania State University
University Park, PA

Emeritus Senior Research Scientist
Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, MI

WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

This Page Intentionally Left Blank

Margins of Error

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

Margins of Error

A Study of Reliability in Survey Measurement

Duane F. Alwin
McCourtney Professor of Sociology and Demography
Pennsylvania State University
University Park, PA

Emeritus Senior Research Scientist
Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, MI

WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Alwin, Duane F. (Duane Francis), 1944–
Margins of error : a study of reliability in survey measurement / Duane F. Alwin.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-08148-8 (cloth : acid-free paper)
1. Surveys. 2. Error analysis (Mathematics) I. Title.
HA31.2.A5185 2007
001.4'33—dc22 2006053191

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Dedicated to

Edgar F. Borgatta

and to the memory of

T. Anne Cleary
David J. Jackson
Charles F. Cannell

This Page Intentionally Left Blank

Contents

Preface

Among social scientists, almost everyone agrees that without valid measurement, there may be little for social science to contribute in the way of scientific knowledge. A corollary to this principle is that *reliability of measurement* (as distinct from validity) is a necessary, although not sufficient, condition for valid measurement. The logical conclusion of this line of thinking is that whatever else our measures may aspire to tell us, it is essential that they are reliable; otherwise they will be of limited value. There is a mathematical proof of this assertion (see Chapters 3 and 12), but the logic that underlies these ideas is normally accepted without formal proof: If our measures are *unreliable*, they cannot be trusted to detect patterns and relationships among variables of interest. Thus, reliability of measurement is a sine qua non of any empirical science.

To be somewhat more concrete, there are several reasons to be concerned with the existence and consequences of errors in social measurement. First and foremost, if we are aware of the processes that generate measurement error, we can potentially understand the nature of our results. A presumptive alternative interpretation for any research result is that there are methodological errors in the data collection, and we must rule out such methodological artifacts as explanatory variables whenever we draw inferences about differences in patterns and processes. Second, if we know about the nature and extent of measurement errors, we may (in theory) get them under better control. In the second chapter of this book, I "deconstruct" the data gathering process in survey research into *six major elements* of the response process—question adequacy, comprehension, accessibility, retrieval, motivation, and communication—and argue that discerning how measurement errors result from these components helps researchers reduce errors at the point where they are most likely to occur. Third, measurement errors affect our statistical inferences. Measurement unreliability inflates estimates of population variance in variables of interest and, in turn, biases estimates of standard errors of population means and other quantities of interest, inflating confidence intervals. Statistical analyses that ignore unreliability of variables underestimate the strength and significance of the statistical association between those variables. This underestimation not only makes the results of such analyses more conservative from a scientific perspective; it also

increases the probability of type II error and the consequent rejection of correct, scientifically productive hypotheses about the phenomena of interest. Even in the simplest regression models, measurement unreliability in predictor variables generally biases regression coefficients downward, making it more difficult to reject the null hypothesis; and unreliability in both dependent and independent variables attenuates estimates of statistical associations. With appropriate measurement designs, it is possible to isolate some types of errors statistically and control for them in the analysis of data.

Over the past two decades (i.e., since the mid-1980s), we have seen a burgeoning research literature on survey methodology, focusing especially on problems of measurement. Indeed, volumes have been written about measurement errors. We probably have a better than ever understanding about the sources of measurement errors, particularly those involving cognitive processes and the effects of question wording. But little effort has been undertaken to quantify the extent of unreliability in the types of measures typically used in population surveys to help us assess the extent of its biasing effects. To be blunt, our knowledge about the nature and extent of measurement errors in surveys is meager, and our level of understanding of the factors linked to the design of questions and questionnaires that contribute to their presence is insufficient. Errors of measurement—the general class of phenomena of which unreliability is a particular type—are a bit like what Mark Twain reportedly said about the weather: "Everybody talks about the subject, but nobody does anything about it."

In this book, I argue that considering the presence and extent of measurement errors in survey data will ultimately lead to improvements in data collection and analysis. A key purpose of studies of measurement errors is to identify which types of questions, questionnaires, and interviewer practices produce the most valid and reliable data. In the chapters that follow, I consider ways in which the extent of measurement errors can be detected and estimated in research in order to better understand their consequences. The major vehicle for achieving these purposes involves a study of nearly 500 survey measures obtained in surveys conducted at the University of Michigan over the past two or three decades. Assembling information on reliability from these data sources can help improve knowledge about the strengths and weaknesses of survey data. The results of this research should be relevant to the general tasks of uncovering the sources of survey measurement error and improving survey data collection through the application of this knowledge.

Although information about the level of reporting reliability in the standard survey interview is lacking, a small and growing cadre of investigators is addressing this issue. Given the substantial social and economic resources invested each year in data collection to satisfy social and scientific information needs, questions concerning the quality of survey data are strongly justified. Without accurate and consistent measurement, the statistical tabulation and quantitative analysis of survey data hardly makes sense. Knowledge has only recently been cumulating regarding the factors linked to the quality of measurement, and I hope this study will contribute to this body of work.

Acknowledgments

There are many people to whom credit for the success of this project needs to be acknowledged. I am indebted to my daughters—Heidi, Abby, and Becky—and their families, for their love and support. My wife, Linda Wray, a social scientist in her own right, understands how important this project has been to me and has given me an unending amount of encouragement. My deepest gratitude goes to Linda—none of this would have been possible without your help, encouragement, and love (not to mention the copyediting). And more than anything, Linda, your confidence in me made all the difference.

The contributions of my research assistants over the years to this project have been indispensable. I can honestly say that this book would have been finished much earlier if it were not for my research assistants—Ryan McCammon, Dave Klingel, Tim Manning, Frank Mierzwa, Halimah Hassan, and Jacob Felson—but it would not have been as good. The efforts of Dave and Ryan in particular, who insisted that I "get it right" (to the extent that is ever possible) slowed me down. I recall many occasions on which they insisted that we not leave as many "stones unturned," which often encouraged me to take the analysis further, or to consider an alternative set of estimates, or even in some cases actually recode the data, in order to make the study as good as it could be. Ryan McCammon has played an extraordinarily important role over the past few years in helping me bring this project to completion. His technical knowledge, his expert advice in statistical matters, and his understanding of measurement issues have helped forge our collaboration on these and other projects. Pauline Mitchell, administrative assistant and word-processing expert par excellence, made it possible to express our results in numeric and graphic form. Pauline is responsible for the more than 120 tables, figures, and charts presented here, and I acknowledge her dedicated assistance. I also wish to acknowledge Pauline, Ryan, Linda, and Jake for various copyediting tasks that required them to read portions of earlier drafts. Matthew Williams of Columbus, Ohio provided stellar assistance in producing the camera-ready text for this book. Last but not least, the assistance of John Wiley's senior editor, Steve Quigley, and his staff, is gratefully acknowledged.

Some chapters presented here build on other previously published writings. Part of Chapter 3 is related to my recent entry titled "Reliability" in the *Encyclopedia of*

Social Measurement, edited by K. Kempf-Leonard and others (2004). Portions of this work are reprinted with permission of Elsevier Ltd, UK. Parts of Chapter 4 are related to my earlier publications dealing with the multitrait-multimethod approach, specifically my 1974 chapter, "Approaches to the interpretation of relationships in the multitrait-multimethod matrix" in *Sociological Methodology 1973–74*, edited by H.L. Costner (San Francisco: Jossey-Bass), and my 1997 paper, "Feeling thermometers vs. seven-point scales: Which are better?," which appeared in *Sociological Methods and Research* (vol. 25, pp. 318–340). And finally, the material presented in Chapter 10 is an extension of a chapter, "Aging and errors of measurement: Implications for the study of life-span development," in *Cognition, Aging, and Self-Reports*, edited by N. Schwarz, D. Park, B. Knäuper, and S. Sudman (Philadelphia: Psychology Press, 1999).

In 1979, I took a job at the Survey Research Center of the Institute for Social Research at the University of Michigan, where I worked for nearly 25 years. This move was motivated in part by a desire to learn more about survey research, and it paid off. At Michigan I learned how surveys were actually conducted and how measurement errors were created—by participating in surveys I conducted, as well as those carried out by others (more than a dozen surveys in all). Also, I learned how questionnaires were designed, interviewers were trained, pretesting of questionnaires was carried out, and the data were actually gathered in the field. Although it is not possible to credit all the sources of my education at Michigan, I wish to acknowledge particularly the support over the years of Jon Krosnick, Charles Cannell, Leslie Kish, Arland Thornton, Bob Groves, Jim House, Regula Herzog, Norbert Schwarz, and Bill Rodgers. Paul Beatty, whose dissertation at Michigan (Beatty, 2003) I was fortunate to supervise, taught me a great deal about the importance of cognitive factors in responses to surveys.

Over the years this project received support from the National Science Foundation and the National Institute on Aging. During 1992–1995 the project was supported by an NIA grant, "Aging and errors of measurement" (R01-AG09747), and this led to a current NIA-funded project, "Aging and the reliability of measurement" (R01-AG020673). These projects were instrumental in the work reported in Chapters 10 and 11. In between these two projects, I received support from the NSF for the project, "The reliability of survey data" (SES-9710403), which provided the overall rationale for assembling a database containing estimates of reliability and question characteristics.

Finally, I also acknowledge the support of the Tracy Winfree and Ted H. McCourtney Professorship, the Population Research Institute, and of the College of the Liberal Arts at Pennsylvania State University, support that allowed me some additional time to complete this research.

Foreword

Some projects take a long time to complete—this is true (for better or worse) of the present one. In a very real sense, the idea for this project began nearly 40 years ago, when I was a graduate student in sociology at the University of Wisconsin. In 1968, I was taking courses in reliability theory, factor analysis, and item response theory in the Department of Educational Psychology at Wisconsin (from Anne Cleary, Chester Harris and Frank Baker) in order to fulfill a Ph.D. minor requirement. At the time, I recall wondering if it would be possible to apply some of the ideas from classical psychometric theory to social science data. Large-scale survey studies were beginning to achieve greater popularity as a mainstay for social science research, and I had read several of the famous critiques about survey research. From those seeds of curiosity sown so many years ago, I now have come to understand how those "response" theories of measurement may be fruitfully applied to survey data.

I recall reading early on about the questionable role of surveys in the development of social science. The field of survey research was so under-developed in the 1950s and 1960s that Herbert Blumer (1956) could wage what seemed to many to be a credible attack on "variable analysis" in social science. Interestingly, Blumer's argument focused on the issues of *reliability* and *validity* of survey data. He believed that survey data had a high degree of reliability, but were of questionable validity. As I argue in this book, Blumer may have assumed too much about the reliability of survey data. Certainly there was little attention to the issue at the time he was writing. But, in fact, Quinn McNemar, a psychometrician, had (10 years earlier) written an important review of survey research methods in which he pointed out that survey researchers had largely ignored the problem of reliability, depending without qualms on results from single questions (McNemar, 1946). Psychometric methods had not yet made their way into survey analysis, and it was not known how to incorporate measurement errors into models for the analysis of survey data. Even the most highly regarded proponents of the quantitative analysis of survey data, Robert Merton and Paul Lazarsfeld, admitted that there was very little discussion of the art of analyzing material once it has been collected (Merton and Lazarsfeld, 1950). Later, in 1968, sociologist James Coleman observed that the investigation of response unreliability was an almost totally underdeveloped field, because of the

lack of mathematical models to encompass both unreliability and change (Coleman, 1968). These arguments piqued my interest in problems of survey measurement.

I found these issues to be compelling, issues I wanted to explore further. It was my good fortune to have been accepted into an NIH training program in quantitative methodology during my graduate studies, a program initiated by Edgar Borgatta. Exposure to his work, along with that of a growing field of "sociological methodology," which included the work of James Coleman, Hubert M. Blalock, Jr., Herbert L. Costner, David R. Heise, Robert M. Hauser, and Karl Jöreskog, among others, did much to help develop an understanding of the nature of social measurement (see Blalock, 1965; Costner, 1969; Hauser and Goldberger, 1971). As one who reads the present book will discover (see Chapter 5), Dave Heise's paper on the separation of unreliability and true change in repeated measures designs and Karl Jöreskog's related work on simplex models were critical to the development of this research program (see Heise, 1969; Jöreskog, 1970).

Many of my early publications dealt with these matters (e.g., Alwin 1973, 1974), and these concerns have remained an important focus for a substantial portion of my scholarly work. It is a remarkable thing to have been driven and influenced most of my professional life by a general concern with the quality of data on which the inferences of social scientists are based. And although I have worked on a range of other topics throughout my career, a concern with measurement issues has been a keystone of my work.

I have dedicated this book to those from whom I learned the most about measurement—people whose influence I am cognizant of almost every day of my life. Ed Borgatta, now retired from the University of Washington, is without question one of the most cherished mentors I have ever had—his knowledge of measurement and his hard-nosed approach to modeling social data are attributes I hope I have passed along to my students. Anne Cleary—whose life was taken at a young age by a senseless act of terror—was an extraordinarily talented mentor who taught me just about everything I know about classical measurement theory. David Jackson, a colleague in graduate school, was my best friend. His life was taken by cancer on October 1, 2001, just a few weeks after 9/11. I still feel the grief of losing Dave, but I can say this—Dave taught me so much about measurement that I cannot think about the content of this book without thinking of what I learned from him. Charlie Cannell hired me for the job at the University of Michigan and exposed me to interviewing methodology and survey research in a way I would never have thought possible—I have only the fondest of memories of my contact with Charlie and what I learned from him.

This book took many years to complete, and the research spanned my tenure across several academic institutions. The work was influenced by many esteemed colleagues, friends, and family, and the research was supported by two federal funding agencies. I believe the time it has taken and the influence of others only strengthened the final product. To all those who read this book, I hope the underpinnings of the approach and the importance of the research agenda can have an impact on future research. To my colleagues, friends, and family who made this project possible, you have my deepest appreciation.

DUANE F. ALWIN

State College, Pennsylvania
October 2006

CHAPTER ONE

Measurement Errors in Surveys

> Quality . . . you know what it is, yet you don't know what it is. But that's self-contradictory. But some things are better than others, that is, they have more quality. But when you try to say what the quality is, apart from the things that have it, it all goes poof! . . . But if you can't say what Quality is, how do you know what it is, or how do you know that it even exists? If no one knows what it is, then for all practical purposes it doesn't exist at all. But for all practical purposes it really does exist. . . . Obviously some things are better than others . . . but what's the "betterness"? . . . So round and round you go, spinning mental wheels and nowhere finding any place to get traction. What the hell is Quality? What is it?
>
> Robert M. Pirsig, *Zen and the art of motorcycle maintenance* (1974)

Measurement issues are among the most critical in scientific research because analysis and interpretation of empirical patterns and processes depend ultimately on the ability to develop high quality measures that accurately assess the phenomenon of interest. This may be more difficult in the social and behavioral sciences as the phenomena of interest are often not well specified, and even when they are, the variables of interest are often difficult to observe directly. For example, concepts like religiosity, depression, intelligence, social status, attitudes, psychological well being, functional status, and personality may be difficult to measure precisely because they largely reflect unobserved processes. Even social indicators that are more often thought to directly assess concepts of interest, e.g., variables like education, or income, or race, are not free of specification errors. Clearly, the ability to define *concepts* precisely in a conceptually valid way, the translation of these concepts into *social indicators* that have an empirical referent, and the development of survey *measures* of these indicators all bear on the extent of measurement errors. In addition, measurement problems in social science are also critically related to the nature of the communication and cognitive processes involved in gathering data from respondents (e.g., Bradburn and Danis, 1984; Cannell, Miller and Oksenberg, 1981; Krosnick, 1999; Schwarz, 1999a, 1999b; Sirken, Herrmann, Schechter, Schwarz, Tanur, and Tourangeau, 1999; Sudman, Bradburn and Schwarz, 1996; Tourangeau, 1984; Tourangeau and Rasinski, 1988; Tourangeau, Rips, and Rasinski, 2000).

With its origins in 19th-century Europe and pre-World War II American society, survey research plays an extraordinarily important role in contemporary social sciences throughout the world (Converse, 1987). Vast amounts of survey data are collected for many purposes, including governmental information, public opinion and election surveys, advertising and marketing research, as well as basic social scientific research. Some have even described survey research as the via regia for modern social science (Kaase, 1999, p. 253)—the ideal way of conducting empirical science. Many would disagree with the proposition that surveys are the *only* way to do social science, but there would be hardly any dissent from the view that survey research has become a mainstay for governmental planning, the research of large numbers of academic social scientists, and the livelihoods of growing numbers of pollsters, and marketing and advertising researchers.

1.1 WHY STUDY SURVEY MEASUREMENT ERROR?

The basic purpose of the survey method is to obtain information from a sample of persons or households on matters relevant to researcher or agency objectives. The survey interview is conceived of as a setting in which the question-answer format is used by the researcher to obtain the desired information from a respondent, whether in face-to-face interview situations, via telephone interviews, or in self-administered questionnaires. Many aspects of the information gathering process may represent sources of measurement error: aspects of survey questions; the cognitive mechanisms of information processing and retrieval; the motivational context of the setting that produces the information; and the response framework in which the information is then transmitted (see, e.g., Alwin, 1991b; Alwin, 1992; Krosnick, 1999; Krosnick and Alwin, 1987, 1988, 1989; Schaeffer, 1991b; O'Muircheartaigh, 1997).

Given the substantial social and economic resources invested each year in data collection to satisfy social and scientific information needs, questions concerning the quality of survey data are strongly justified. Without accurate and consistent measurement, the statistical tabulation and quantitative analysis of survey data hardly makes sense; yet there is a general lack of empirical information about these problems and very little available information on the reliability of measurement from large scale population surveys for standard types of survey measures. For all the talk over the past decade or more concerning measurement error (e.g., Groves, 1989, 1991; Biemer, Groves, Lyberg, Mathiowetz and Sudman, 1991; Biemer and Stokes, 1991; Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, and Trewin, 1997), there has been very little empirical attention to the matter. Indeed, one prescient discussion of measurement errors in surveys even stated that "we know of no study using a general population survey that has attempted to estimate the reliabilities of items of the types typically used in survey research" (Bohrnstedt, Mohler, and Müller, 1987, p. 171). Knowledge has only recently been cumulating regarding the factors linked to the quality of measurement, and we hope this study will contribute to this body of work.

Errors occur in virtually all survey measurement, regardless of content, and the factors contributing to differences in unreliability of measurement are worthy of scrutiny. It is well known that statistical analyses ignoring unreliability of measures

generally provide biased estimates of the magnitude and statistical significance of the tests of mean differences and associations among variables. Although the resulting biases tend to underestimate mean differences and the strength of relationships making tests of hypotheses more conservative, they also increase the probability of type II errors and the consequent rejection of correct, scientifically valuable hypotheses about the effects of variables of interest (see Biemer and Trewin, 1997). From a statistical point of view there is hardly any justification for ignoring survey measurement errors.

1.2 SURVEY ERRORS

Terms that are often associated with assessments of survey quality, for example, the terms "bias," "reliability," and "validity," are often used in ambiguous ways. Sometimes they are used very generally to refer to the overall stability or dependability of survey results, including the extent of sampling error, nonresponse bias, instrument bias, as well as reporting accuracy. Other times they are used in a much more delimited way, to refer *only* to specific aspects of measurement error, distinguishing them from assessments of other types of survey errors. It is therefore useful to begin this discussion by clarifying how we might think about various types of survey error, how they differ from one another, and how we might arrive at a more precise definition of some of the terms frequently used to refer to levels of survey data quality involving measurement errors in particular.

In his path-breaking monograph, *Survey errors and survey costs*, Robert Groves (1989, p. vi) presents the following framework for considering *four* different types of survey errors:

> *Coverage error.* Error that results from the failure to include some population elements in the sampling frame or population lists.
>
> *Sampling error.* Error that results from the fact that a subset of the population is used to represent the population rather than the population itself.
>
> *Nonresponse error.* Error that results from the failure to obtain data from all population elements selected into the sample.
>
> *Measurement error.* Error that occurs when the recorded or observed value is different from the true value of the variable.

We consider this to be an exhaustive list, and we argue that any type of survey error can be conceptualized within this framework. The presence of any of these types of survey errors can influence the accuracy of the inferences made from the sample data, and the potential for such errors in the application of survey methods places a high priority on being able to anticipate their effects. In the worst case, errors in even one of these categories may be so great as to invalidate *any* conclusions drawn from the data. In the best case, errors are minimized through efforts aimed at their

reduction and/or efforts taken to minimize their effects on the conclusions drawn, in which cases stronger inferences can be made on the basis of the data.

All of these types of *survey errors* are to some extent present in the data we collect via survey methods. It is important for users of survey data to realize that these various survey errors are *nested* in important ways (see Figure 1.1). To describe this aspect of the phenomenon, we use the metaphor of a set of interrelated structures, each inside the next, like a set of Russian *matrioshka* dolls, in which distinct levels of "nestedness" represent different "compoundings" of error (see Alwin, 1991). *Non-response errors* are nested within *sampling errors*, for example, because only those cases sampled have the opportunity to participate and provide a response to the survey and the cases representing nonresponse or missing cases, depend on which elements of the population are selected into the sample (Groves and Couper, 1998; Groves, Dillman, Eltinge, and Little, 2002). Similarly, *sampling errors* are nested within *coverage errors* because clearly the subset of the population sampled depends on the coverage of the sampling frame. Finally, measurement errors are nested within those cases that have provided a response, although typically we study processes of measurement error as if we were studying those processes operating at the population level. Inferences about measurement error can only be made with the realization that they pertain to respondents from samples of

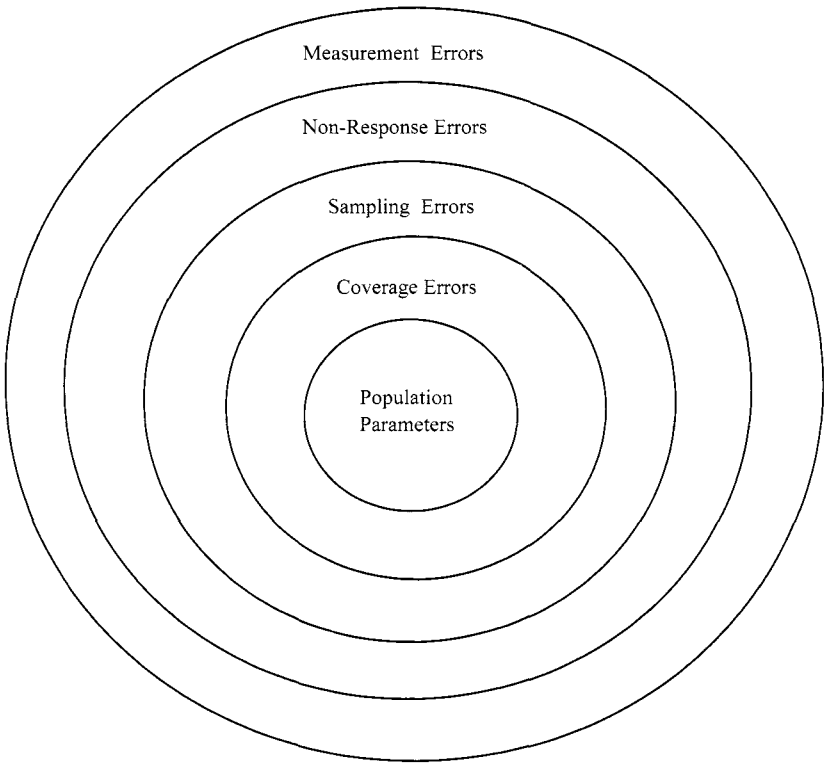

Figure 1.1. The relationship of sources of survey errors.

specified populations, and it is important to realize, thus, that our inferences regarding those processes are constrained by the levels of nestedness described here.

1.2.1 Classifying Types of Survey Error

There are a number of different ways to think about the relationship among the several types of survey error. One way is to describe their relationship through the application of classical statistical treatments of survey errors (see Hansen, Hurwitz, and Madow, 1953). This approach begins with an expression of the mean square error (MSE) for the deviation of the sample estimator ($\bar{y}$) of the mean (for a given sampling design) from the population mean (μ), that is, MSE ($\bar{y}$) = $E(\bar{y} - \mu)^2$. This results in the standard expression:

$$\text{MSE}(\bar{y}) = \text{Bias}^2 + \text{Variance}$$

where *Bias*^{2} refers to the square of the theoretical quantity $\bar{y} - \mu$, and *Variance* refers to the variance of the sample mean $\sigma^2_{\bar{y}}$ Within this statistical tradition of conceptualizing survey errors, *bias* is a *constant source of error* conceptualized at the sample level. *Variance*, on the other hand, represents variable errors, also conceptualized at the sample level, but this quantity is obviously influenced by the within-sample sources of response variance normally attributed to measurement error.

Following Groves' (1989) treatment of these issues, we can regroup coverage, sampling, and nonresponse errors into a category of *nonobservational errors* and also group measurement errors into a category of *observational errors. Observational errors* can be further subclassified according to their sources, e.g., into those that are due to interviewers, respondents, instruments, and modes of observation. Thus, Groves' fourfold classification becomes even more detailed, as seen in Table 1.1. Any treatment of survey errors in social research will benefit from the

Table 1.1. A classification of some types of survey errors

MSE ($\bar{y}$) =	Bias^2	+	Variance
	Nonobservational Errors		*Nonobservational Errors*
	Coverage bias		Coverage area variance
	Sampling bias		Sampling error variance
	Nonresponse bias		Nonresponse error variance
	Observational Errors		*Observational Errors*
	Interviewer bias		Interviewer error variance
	Respondent bias		Respondent error variance
	Instrument bias		Instrument error variance
	Mode bias		Mode error variance

use of this classification, and further, any comparison of results across settings (e.g., across national surveys) will benefit from an understanding of the potential role of these components in the production of similarities and differences observed across settings. Ultimately, while this classification scheme is useful for pinpointing the effects of survey errors on sample estimates of means and their variances, it is also important to understand what (if anything) might be done to estimate these effects and the contributions of error sources to the understanding the results of research studies. This book focuses on one of these types of error—*survey measurement errors*—and it is hoped that the program of research summarized in subsequent chapters will improve our understanding of the effects of measurement errors on the results of surveys.

1.3 SURVEY MEASUREMENT ERRORS

Measurement represents the link between theory and the analysis of empirical data. Consequently, the relationship between measures of empirical indicators and the theoretical constructs they represent is an especially important aspect of measurement, in that ultimately the inferences drawn from the empirical data are made with respect to more abstract concepts and theories, not simply observable variables. Measurement, thus, requires the clear specification of relations between theoretic constructs and observable indicators. In addition, obtaining "measures" of these indicators involves many practical issues, including the specification of questions that operationalize the measures, and in the case of survey research the processes of gathering information from respondents and/or households.

As I indicated at the beginning of this chapter, the specification of the linkage between theory and measurement is often viewed as more difficult in the social and behavioral sciences, as the phenomena of interest are often not very well specified, and even where they are, the variables are often difficult or impossible to observe directly. The diagram in Figure 1.2 illustrates the fundamental nature of the problem of measurement. Here I have adopted a *three-ply distinction* between constructs, indicators, and measures, depicting their interrelationships. *Constructs* are the theoretical variables referred to in theoretical or policy discussions about which information is desired. *Indicators* are the empirical referents to theoretical constructs. In social surveys *measures* consist of the question or questions that are used to obtain information about the indicators. The layered nature of the distinctions of interest here can be illustrated with an example. Socioeconomic status is an example of a theoretical construct, derived from sociological theory, which can be indexed via any number of different social indicators, e.g., education, occupation, income level, property ownership. Normally, one considers such indicators as imperfect indicators of the theoretical construct, and often researchers solve this problem through the use of *multiple indicators*, combining different indicators using MIMC (multiple-indicator multiple-cause) models or common factor models for analysis (see Alwin, 1988).

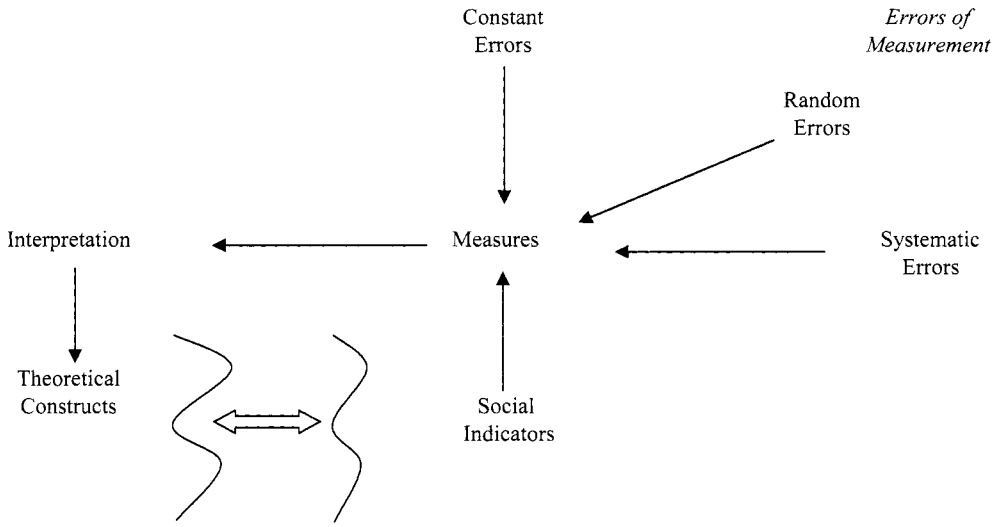

Figure 1.2. The relationship between constructs, indicators, measures, and measurement errors.

It is important in this context not to confuse the indicators of concepts with the theoretical constructs they are thought to reflect. To do so loses sight of the purpose of measurement. In principle, it is the theoretical constructs that implicate particular indicators, not the reverse. In Kuhn's (1961, p. 190) words: "To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly." The linkage between scientific theories and scientific measurement is therefore rarely traced backward, specifying constructs entirely in terms of what can be assessed at the operational level. Still, it is clearly possible for scientific data to be gathered in service of theories that are wrong or ill-conceived, and in some cases new theoretical constructs may be needed in order to account for the empirical data. The relation between the two therefore should probably be conceived of as reciprocal (as depicted in Figure 1.2).

It is also important to realize that given a particular indicator, there may be *multiple measures* that can be used to assess variation in the indicator. *It is crucial in the present context that a distinction be maintained between the idea of multiple indicators and that of multiple measures, as they refer to very different things.* In the example of the indicator, "level of education," measures may include such things as questions focusing on the number of years of schooling completed, or levels of certification, or even a test of knowledge gained from school. All such things may be legitimate measures of education, even though they may assess ostensibly different things, and it may often be the case that one is more appropriate in a given cultural context than another. The point is that within the survey context a "measure" relies on a question or a set of questions that provide the information needed to construct the indicator, and therefore "multiple" measures involve multiple replications of the measure of a given indicator.

On a practical level, once concepts and indicators are defined and agreed on, measurement is possible only if we can assume some type of equivalence across units of observation, e.g., respondents or households. As Abraham Kaplan (1964) has pointed out, the essence of measurement at an operational level lies in the principle of *standardization*, that is, the principle that units of magnitude have constancy across time and place. This, of course, implies that a regularity or stability of measures is required in order for there to be valid comparisons made across units of observation. The equivalence of some units of measurement may seem natural, as in the physical sciences—measures such as weight, mass, distance, or time—but in the social sciences most metrics are completely arbitrary, and standardization is often more an objective than a reality. However, without the ability or willingness to make such strong assumptions about *standardization* of measurement, the comparative usefulness of the information collected in surveys may be in doubt.

Efforts to measure using the survey method involve a two-way process of communication that *first* conveys what information is needed from the respondent to satisfy the objectives of the research and *second* the transmission of that information back to the researcher, usually, but not exclusively, via interviewers. The central focus of measurement in surveys, then, involves not only the specification of a *nomological network* of concepts and their linkage to "observables," but also a focus on the processes of gathering information on "measures."

A concern with the existence and consequences of errors made in this "two step" process motivates their consideration, and hence such consideration can play a role in reducing errors and better understanding the answers to survey questions. This chapter and subsequent ones focus specifically on the conceptualization and estimation of the nature and extent of *measurement error* in surveys and establish the rationale for the consideration of measurement errors when designing survey research. The diagram in Figure 1.2 specifies three types of measurement errors—*constant errors*, *random errors*, and *systematic errors*—that are of concern to researchers who have studied response errors in surveys. This and subsequent chapters will provide a more complete understanding of both the nature and sources of these types of error. We begin with a discussion of the *standards* used in the evaluation of the quality of survey measurement.

1.4 STANDARDS OF MEASUREMENT

Writing on "scales of measurement," Otis Dudley Duncan (1984a, p. 119) observed: "measurement is one of many human achievements and practices that grew up and came to be taken for granted before anyone thought to ask how and why they work." Thus, we argue that one of the biggest challenges for survey research is to figure out "how and why" survey measurement works, but also to assess when it does not work. One of the greatest potential impediments to the meaningful analysis of survey data is the existence of imperfect measurement; imperfect either in providing a *valid* correspondence of indicators to the target concept(s) of interest, or in producing *reliable* measures of those indicators.

In order to *measure* a given quantity of interest, it is necessary to (1) specify certain rules of correspondence between concepts and indicators, (2) establish the nature of the dimensionality inherent in the phenomenon, (3) choose a particular metric in which to express variation in the dimension or dimensions of the concept being measured, and (4) specify the necessary operational procedures for gathering information that will reveal the nature of differences in the phenomenon. As noted earlier, the term "measurement" implies equivalence. In *The conduct of inquiry*, Abraham Kaplan (1964, pp. 173–174) observed: "Measurement, in a word is a device for *standardization*, by which we are assured of the equivalences among objects of diverse origin. This is the sense that is uppermost in using phrases like 'a measure of grain': measurement allows us to know what quantity we are getting, and to get and give just what is called for." *Equivalence*, then, across all of the elements mentioned above, is the key to measurement. It should come as no surprise, then, that one of the major criticisms of quantitative approaches is the comparability of units and therefore of responses across respondents.

In this and subsequent chapters I discuss the issue of obtaining useful information in surveys and the problem of assessing the extent of *measurement error* and the factors that contribute to such errors. As I point out in the next chapter, errors of measurement can intrude at many different points in the gathering of survey data, from the initial *comprehension* of the question by the respondent, to the *cognitive processing* necessary to access the requested information, through to the production of a response. Clearly, the magnitude of such errors qualify the meaning one can attach to the data, and ultimately the confidence we place in survey research strategies depends intimately on the extent of measurement errors in the data.

1.5 RELIABILITY OF MEASUREMENT

Let us return to the above definition of measurement error as *the difference between the recorded or observed value and the true value of the variable* (see Groves, 1989, 1991). There have been two basic approaches to minimizing this type of error. The first is to emphasize the reduction of errors in the collection of survey data through improved techniques of questionnaire design, interviewer training and survey implementation. The second is to accept the fact that measurement errors are bound to occur, even after doing everything that is in one's power to minimize them, and to model the behavior of errors using statistical designs. The tradition in psychology of "correcting for attenuation" is an example of an approach that adjusts sample estimates of correlations based on available information about the reliabilities of the variables involved (Lord and Novick, 1968). More recently, structural equation models (or LISREL-type models) used to model response errors in surveys are another example of such an approach (see Alwin and Jackson, 1979; Bollen, 1989).

Earlier we noted that sometimes the term *reliability* is used very generally to refer to the overall stability or dependability of research results, including the absence of population specification errors, sampling error, nonresponse bias, as well as various forms of measurement errors. Here (and throughout the remainder of this book) we

use the term in its more narrow *psychometric* meaning, focusing specifically on the absence of measurement errors. Even then, there are at least two different conceptions of error—random and nonrandom (or systematic) errors of measurement—that have consequences for research findings. Within the psychometric tradition the concept of reliability refers to the absence of *random error.* This conceptualization of error may be far too narrow for many research purposes, where reliability is better understood as the more general absence of measurement error. However, it is possible to address the question of reliability separately from the more general issue of measurement error and in subsequent chapters I point out the relationship between random and nonrandom components of error.

Traditionally, most attention to reliability in survey research is devoted to item analysis and scale construction [e.g., calculation of Cronbach's (1951) alpha (α)], although including *multiple indicators* using SEM models or related approaches is increasingly common (Bollen, 1989). While these procedures are invaluable and likely to reduce the impact of measurement errors on substantive inferences, they have not informed survey researchers of the nature and sources of the errors of concern. Further, these approaches generally cannot address questions of reliability of survey questions because they focus on composite scales or on common factor models of multiple indicators (rather than multiple measures). It is well known that quantities like Cronbach's α depend on factors other than the reliabilities of the component items.

While some attention has been given to this issue, we still know very little about patterns of reliability for most types of survey measures. Increasing information on survey data reliability may improve survey data collection and its analysis, and estimates of the reliability of survey measures can help researchers adjust their models. There is a large body of literature in statistics that deals with the problems of conceptualizing and estimating measurement errors (e.g., Biemer and Stokes,1991; Groves, 1991; Lyberg et al., 1997). Until fairly recently, however, little attention was paid to obtaining empirical estimates of measurement error structures (see, e.g., Alwin and Jackson, 1979; Alwin, 1989, 1992, 1997; Alwin and Krosnick, 1991b; Andrews, 1984; Bielby and Hauser, 1977; Bielby et al., 1977a, 1977b; Bound et al., 1990; Duncan et al., 1985; McClendon and Alwin, 1993; Rodgers, Andrews, and Herzog, 1992; Saris and Andrews, 1991; Saris and van Meurs, 1990; Scherpenzeel, 1995; Scherpenzeel and Saris, 1997).

1.6 THE NEED FOR FURTHER RESEARCH

Despite increasing attention to problems of measurement error in survey design and analysis, there are three basic problems that need to be addressed: (1) the lack of attention to measurement error in developing statistical modeling strategies, (2) the relative absence of good estimates of reliability to adjust for measurement error, and (3) the lack of information about how measurement error varies across subgroups of the population, for example, by age and levels of education. On the first point, many multivariate analysis techniques common in analysis of survey data—e.g., hierarchical linear models (HLM) and event history models (EHM)—have ignored explicit

consideration of problems of measurement error. On the whole these approaches have not incorporated psychometric adjustments to the model (see, e.g., Bryk and Raudenbush, 1992; Tuma and Hannan, 1984; Petersen, 1993). Of course, there are exceptions in the area of event history models (e.g., Holt, McDonald and Skinner, 1991) and multilevel models (e.g., Goldstein, 1995). In fact, Goldstein devotes an entire chapter to applying estimates of reliability to multilevel models. It is important to note that rather than being a product of the HLM modeling strategy, reliability information is assumed to exist. Goldstein (1995, p. 142) states that in order for such models to adjust for measurement error, one must "assume that the variances and covariances of the measurement errors are known, or rather that suitable estimates exist."

By contrast, within the structural equation models (SEM) tradition, there has ostensibly been considerable attention to the operation of measurement errors. It is often stated that LISREL models involving multiple indicators "correct for measurement error." There are some ways in which this is true, for example, when analysts employ "multiple measures" (the same measure repeated either in the same survey or in repeated surveys) (e.g. Bielby, Hauser and Featherman, 1977a, 1977b; Bielby and Hauser, 1977; Hauser, Tsai, and Sewell, 1983; Alwin and Thornton, 1984). The same conclusion does not generalize to the case where "multiple indicators" (within the same domain, but not identical measures) are employed. There is, unfortunately, considerable confusion on this issue (see Bollen, 1989), and in subsequent chapters (see especially Chapter 3) I develop a clarification of the critical differences between "multiple measures" and "multiple indicators" approaches and their differential suitability for the estimation of reliability.

With regard to *the absence of reliability estimates*, current information is meager and unsystematic, and there are several problems associated with obtaining worthwhile estimates of measurement quality. Empirical research has not kept pace with the theoretical development of statistical models for measurement error, and so, while there are isolated studies of the behavior of measurement error, there has been no widespread adoption of a strategy to develop a database of reliability estimates. On the basis of what we know, we must conclude that regardless of how valid the indicators we use and no matter how rigorously the data are collected, *survey responses are to some extent unreliable.* More information needs to be collected on the relative accuracy of survey data of a wide variety of types (e.g., facts, attitudes, beliefs, self-appraisals) as well as potential sources of measurement error, including both respondent characteristics (e.g., age, education) and formal attributes of questions.

1.7 THE PLAN OF THIS BOOK

In this chapter I have stressed the fact that whenever measures of indicators are obtained, errors of measurement are inevitable. I have argued that one of the most basic issues for consideration in survey research is that of measurement error. This is of critical importance because measurement requires the clear specification of relations between theoretical constructs and observable indicators, as well as the

specification of relations between observable indicators and potential measures. In the next chapter I "deconstruct" the data gathering process into its components in order to recognize the considerable potential for measurement error at each step in the reporting process. That chapter and subsequent ones focus specifically on the conceptualization and estimation of the nature and extent of *measurement error* in surveys and establish the rationale for the consideration of potential measurement errors when designing and conducting survey research.

There are several reasons to be concerned with the existence and consequences of errors made in survey measurement. First and foremost, *an awareness of the processes that generate measurement error* can potentially help us understand the nature of survey results. One of the presumptive alternative interpretations for any research result is always that there are methodological errors in the collection of data, and thus, it is important to rule out such methodological artifacts as explanatory variables whenever one entertains inferences about differences in patterns and processes. Second, with *knowledge of the nature and extent of measurement errors*, it is possible in theory to get them under better control. Awareness of the *six major elements* of the response process discussed in Chapter 2—question adequacy, comprehension, accessibility, retrieval, motivation, and communication—is important for researchers to understand in order to reduce errors at the point where they are likely to occur. In addition, *with appropriate measurement designs, it is possible to isolate some types of errors statistically* and therefore control for them in the analysis of data.

In the subsequent chapters I argue that the consideration of the presence and extent of measurement errors in survey data will ultimately lead to improvement in the overall collection and analysis of survey data. One of the main purposes of studies of measurement errors is to be able to identify, for example, which types of questions and which types of interviewer practices produce the most valid and reliable data. In the following I consider ways in which the extent of measurement errors can be detected and estimated in research in order to better understand their consequences. The major vehicle for achieving these purposes involves the presentation of results from an extensive National Science Foundation and National Institute of Aging-supported study of nearly 500 survey measures obtained in surveys conducted at the University of Michigan over the past several years. Assembling information on reliability from these data sources can help improve knowledge about the strengths and weaknesses of survey data. It is expected that the results of this research will be relevant to the general task of uncovering the sources of measurement error in surveys and the improvement of methods of survey data collection through the application of this knowledge.

The research addresses the following sets of questions:

- How reliable are standard types of survey measures in general use by the research community?
- Does reliability of measurement depend on the nature of the content being measured? Specifically, is factual information gathered more precisely than attitudinal and/or other subjective data? Also, do types of nonfactual questions (attitudes, beliefs and self-assessments) differ in reliability?

- Does reliability of measurement vary as a function of the source of the information? In the case of factual data, are proxy reports as reliable as self-reports? How reliable are interviewer observations?
- Is reliability of measurement affected by the context in which the questions are framed? Specifically, does the location of the question in the questionnaire, or the use of series or batteries of questions produce detectable differences in levels of measurement error?
- Do the formal properties of survey questions affect the quality of data that results from their use? For example, in attitude measurement, how is reliability affected by the form of the question, the length of the question, the number of response categories, the extent of verbal labeling, and other properties of the response format?
- Are measurement errors linked to attributes of the respondent population? Specifically, how are education and age related to reliability of measurement?

The present research investigates these questions within the framework of a set of working hypotheses derived from theory and research experience on the sources of measurement errors. Simply put, the analysis will focus on explaining variation in reliability due to these several factors.

While a major purpose of this book is to present the empirical results of this study, the goals of this project are more general. In addition to presenting the results of this study, we also review the major approaches to estimating measurement reliability using survey data and offer a critique of those approaches. In Chapter 3 I focus mainly on how repeated measures are used in social research to estimate the reliability of measurement for continuous latent variables. This chapter includes a rigorous definition of the basic concepts and major results involved in classical reliability theory, the major research designs for reliability estimation, methods of internal consistency reliability estimation for linear composites, and recently developed methods for estimating the reliability of single variables or items, including a brief discussion of reliability estimation where the latent variables are latent classes.

This discussion ends with a critique of the standard methods of reliability estimation in common use in survey research—internal consistency estimates—and argues that for purposes of improving survey measurement a focus on the reliability of single measures is necessary. In keeping with this critique, I then review several important developments for the examination of the reliability of single measures: the use of confirmatory factor analysis for the analysis of response errors, including the use of similar methods involving the multitrait-multimethod measurement design, reviewed in Chapter 4, and quasi-simplex models for longitudinal measurement designs, covered in Chapter 5.

Chapter 6 presents the methods used in the present study, including a description of the samples, available measures, statistical designs for reliability estimation, and the problem of attrition in longitudinal designs. The main empirical contribution of this research involves the results of a project whose aim was to assemble a database for survey questions, consisting of question-level information on reliability and question characteristics for nearly 500 variables from large-scale longitudinal surveys of

national populations in the United States. The objective was to assemble information on measurement reliability from several representative longitudinal surveys, not only as an innovative effort to improve knowledge about the strengths and weaknesses of particular forms of survey measurement but also to lay the groundwork for developing a large-scale database on survey measurement reliability that can address basic issues of data quality in the social, economic, and behavioral sciences. This chapter presents these methods in four parts: (1) I describe the longitudinal data sets selected for inclusion in the present analysis, (2) I summarize the measures available in these studies and the conceptual domains represented, (3) I discuss the variety of statistical estimation strategies available for application here, and (4) the problem of panel attrition and its consequences for reliability estimation are addressed.

Using these data, there are three main empirical chapters devoted to the analysis of the contributions of various features of survey questions and questionnaires to levels of measurement unreliability, organized primarily around the topics of *question content, question context,* and the *formal properties of questions.* Chapter 7 discusses the potential effects of topic and source of survey reports on the reliability of measurement and presents results relevant to these issues. Chapter 8 discusses the *architecture of survey questionnaires* and the impact of several pertinent features of questionnaire organization on the reliability of measurement. Chapter 9 presents the basic empirical findings regarding the role of question characteristics in the reliability of measurement.

Assembling information on measurement reliability from these panel surveys will not only improve knowledge about strengths and weaknesses of particular forms of survey measurement but also lay the groundwork for developing a large-scale database on survey measurement reliability that can address basic issues of data quality across subgroups of the population. Chapter 10 presents data on the relationship of respondent characteristics—education and age—to the reliability of measurement. I partition the data by these factors and present reliability estimates for these groups. The most serious challenge to obtaining reasonable estimates of age differences in reporting reliability is the confounding of age with cohort factors in survey data. If cohort experiences were not important for the development of cognitive functioning, there would be no reason for concern. However, there are clear differences among age groups in aspects of experience that are relevant for survey response. Specifically, several studies report that educational attainment is positively related to memory performance and reliability of measurement. Since age groups differ systematically in their amount of schooling attained, cohort factors may contribute spuriously to the empirical relationship between age and measurement errors. In Chapter 11 I introduce several approaches to the study of reliability of measures of categorical latent variables. Finally, I wrap up the presentation of the results of this project by reviewing several topics where future research can profitably focus attention, turning in Chapter 12 to some neglected matters. There I also sketch out some avenues for future research on measurement errors in surveys.

CHAPTER TWO

Sources of Survey Measurement Error

> Get your facts first, and then you can distort them as much as you please.
>
> Mark Twain, quoted by Rudyard Kipling in *From sea to sea and other sketches* (1899)

Reliability of survey measurement has to do with the quality of the information gathered in responses to survey questions. As I clarify in the next chapter, reliability is conceptualized in terms of the "consistency" or "repeatability" of measurement. If one could erase the respondent's memory and instantaneously repeat a particular question, a reliable measure would be one that produced the same response upon repeated measurement. There are a number of different types of measurement error—the principal ones of which are "random" and "systematic" errors—and each has a special relation to reliability of measurement. The problem in quantifying the nature of measurement error in surveys is that errors result from processes that are *unobserved*, and rarely can one directly assess the errors that occur in our data. We are, thus, forced by the nature of these circumstances to construct "models" of how measurement errors happen and then evaluate these models with respect to their ability to represent the patterns that we observe empirically. These models permit us to obtain "estimates" of reliability of measurement, which under optimal circumstances can help us understand the nature of measurement errors. The next chapter (Chapter 3) focuses on understanding how we can estimate reliability of survey measurement. In subsequent chapters I discuss the key strategies that can be used to produce estimates that are interpretable.

There are a number of misconceptions about the reliability of measurement and its utility in evaluating the quality of survey measurement. One argument that is sometimes advanced among survey methodologists is that "consistency of measurement" is of little concern, since what we desire are indications of the "validity of measurement." After absorbing the material in Chapter 3 and in subsequent chapters, one will hopefully be able to integrate the concept of reliability with other aspects of survey quality. On the basis of this understanding of the relationship between the concepts of reliability and validity, one will be able to see the truth to the claim that *reliability* is a necessary condition for validity of measurement and for scientific

inference of any type. Reliability is not a sufficient condition for validity, but it is necessary, and without reliable measurement, there can be no hope of developing scientific knowledge. The obverse of this logic is that if our measures are *unreliable* they are of little use to us in detecting patterns and relationships among variables of interest. Reliability of measurement is therefore a sine qua non of any empirical science.

The researcher interested in understanding the reliability of measurement needs to have three things in order to make progress toward this goal. The first is *a model that specifies the linkage between true and observed variables.* As we noted in the previous chapter, survey methodologists define *measurement error* as error that occurs when the recorded or observed value is different from the true value of the variable (Groves, 1989). An inescapable counterpart to defining measurement error in this way is that it implies that *a definition* of "true value" exists. This *does not mean that a true value is assumed to exist,* only that the model defining "measurement error" also provides a definition of "true value." Also, as we pointed out above, the material we cover in Chapter 3 provides such a set of definitions and a specification of the linkage between true variables, observed variables and measurement errors. The second requirement for estimating reliability of measurement is a *statistical research design* that permits the estimation of the parameters of such a "measurement model," which allows an interpretation of the parameters involved that is consistent with the concept of reliability. Chapters 4 and 5 provide an introduction to two strategies for developing research designs that permit the estimation of measurement reliability for survey questions—cross-sectional designs involving the multitrait-multimethod (MTMM) approach and the quasi-simplex approach using repeated measures in longitudinal designs. Finally, researchers interested in estimating the reliability of measurement in survey data need to have *access to data gathered within the framework of such research designs for populations of interest,* and the empirical study reported in this book (see Chapters 6–10) illustrates how this can be done. Ultimately, as shown in Chapter 3, reliability is a property of specific populations, and while we tend to think about measurement quality and its components (reliability and validity) to be aspects of our measuring instruments, in reality they are population parameters. Although we estimate models for observed variables defined for individuals, we cannot estimate the errors produced by individuals—only the attributes of these processes in the aggregate. Still, understanding what these population estimates mean requires us to understand the processes that generate measurement errors (see O'Muircheartaigh, 1997; Schaeffer, 1991b; Tourangeau, Rips and Rasinski, 2000), and in order to lay the groundwork for thinking about measurement errors and eventually formulating models that depict the nature of these errors, in this chapter we discuss the sources of survey measurement errors.

2.1 THE UBIQUITY OF MEASUREMENT ERRORS

Because of their inherently *unobserved* nature, one normally does not recognize measurement errors when they occur. If we were able to identify errors at the time

they occur and correct them, this would be a wonderful state of affairs, but that is extremely rare. In contrast, measurement errors typically occur unbeknownst to the investigator. For the most part, they result from inadvertent and unintentional human actions, over which we do not necessarily have control—that is, at least within the limits of our best practices as survey researchers. Some respondents or informants obviously do intentionally mislead and knowingly fabricate the information they provide, but such phenomena probably represent a very small percentage of the measurement errors of concern here. Even where respondents are "not telling the truth," they may, in fact, be unaware that they are doing so.

Some examples of what we mean by such measurement errors may be useful. One of the most common sources of error results from the fact that respondents do not understand the question or find the question ambiguous, and are unwilling to come to terms with their lack of comprehension. In our subsequent discussion we put a great deal of emphasis on the importance of writing survey questions that are simple and clear so that they are comprehensible to the respondent. Still, many errors occur because the information sought by the question is unclear to the respondent and s/he either perceives (and then answers) a different question or simply guesses about what the interviewer wants to know. Because of the pressures of social conformity respondents are often motivated to present themselves as knowledgeable, even when they are not, and *this may vary by culture*. Furthermore, survey researchers often discourage respondents from saying "Don't Know." Converse (1976–77) suggested that when respondents say they "Don't Know," this tends to be a function of a lack of information rather than ambivalence or indifference. If respondents do not have the information requested but feel pressure to provide a response, then guessing or fabricating a response will most likely produce an error. This violates the assumption often made in survey research that respondents know the answer to the question and are able to report it.

One famous example of how respondents are often willing to answer questions they do not understand or know little about is provided by Schuman and Presser (1981, pp. 148–149). In an unfiltered form of the question, they asked if respondents favored or opposed the Agricultural Trade Act of 1978 (a fictitious piece of legislation), and some 30.8% of the sample had an opinion about this. In a separate ballot they found that only 10% stated an opinion, when they explicitly offered a Don't Know alternative. It is an interesting question whether this tendency to appear knowledgeable varies by country or cultural context.

Another example of the occurrence of measurement error involves retrospective reporting where the assumption that people have access to distant events is not met. There are few surveys that do not ask respondents something about the past, and perhaps most typically about their own lives, but sometimes the gathering of such information can be alarmingly imprecise (Schwarz and Sudman, 1994). Among other things, researchers routinely question people concerning their social background or early life history, that is, what were some of the characteristics of their family of origin (e.g., father's and mother's occupations, educational levels, native origins, marital status). There are several existing tales of imperfection in the measurement of father's occupation, for example, a variable that is crucial to sociological studies of

social mobility. Blau and Duncan (1967, pp. 14–15) present a somewhat disturbing case. In brief, in a small sample of men for whom they had both census records and survey data, when they compared the child's report of the father's occupation with matched census records, they found only 70% agreement (out of 173 cases). They report:

> Although 30 percent disagreement may appear high, this figure must be interpreted in light of two facts. First, the date of the census and the date at which the respondent attained age 16 could differ by as much as five years; many fathers may actually have changed occupations between the two dates. Second, reinterview studies of the reliability of reports on occupation may find disagreements on the order of 17 to 22 percent . . . even though the information is current, not retrospective.

The Blau and Duncan (1967) testimony on this point is hardly reassuring. They are claiming that the retrospective measurement of father's occupation in the United States Census is not so much worse than contemporaneous self-reports, which may be pretty poor. On the face of it, this seems highly surprising if retrospective reports of parental occupation are as reliable as self-reports, yet the conclusion has been borne out in more recent research reported by Bielby, Hauser, and Featherman (1977a, 1977b) and Hauser, Tsai, and Sewell (1983). A further interesting claim of the Blau and Duncan (1967) report is that recall actually seems to improve with the lapse of time. They report that older men were more likely than younger men to report their father's occupation reliably. This claim, however, should not be pushed too far. It is one thing to suggest that people could accurately recall dates and durations of events and transitions over the life course; it is another to suggest that reliability may increase with age (see Chapter 10).

In other cases researchers are also sometimes interested in people's perceptions of past events and social changes. In the latter type of study there is no way to measure the veridicality of perceptions, unless these reports can be validated by existing data, and it is often found that perceptions of past events are biased (see Alwin, Cohen and Newcomb, 1991). In many cases, it would be a mistake to assume that human memory can access past events and occurrences, and in general, the longer the recall period, the less reliable are retrospections (Dex, 1991). In addition, several factors affect people's abilities to recall the timing of past events. For example, one phenomenon known to exist in retrospective reports is *telescoping*, reporting events as happening more recently than they actually did. Also, more recent experiences may bias people's recollections about their earlier lives, making it difficult, if not impossible, to be certain about their validity.

We often assume that respondents have relatively easy access and can retrieve relatively recent factual material in their own lives. Since most people are employed and in the labor force, we might often assume that they can easily report information regarding their employment and pay. This may not always be the case. Bound, Brown, Duncan, and Rodgers (1990) compared company records on employment and earnings data from a single large manufacturing firm with several thousand employees to responses of a sample of those employees interviewed by telephone. They found that if the questions involve an annual reporting cycle, the measurement

of earnings can be quite good, although not perfect. They find a correlation of .81 between (the log of) the company record of earnings for the previous calendar year and (the log of) the respondent report of their earnings for that year. Even though survey researchers often assume that self reports of "factual" information can be measured quite well, the fact is that even such "hard" variables as income and earnings are not perfectly measured. In this case only two-thirds of the response variance in the survey reports reflects valid variance. Other measures that Bound et al. (1990) examined performed even less well. The correlation between records and survey reports for the pay period immediately preceding the interviews was .46. Similarly a third measure of self-reported earnings from the interview, "usual earnings," correlated .46 with the average value in the records for the preceding 12 weeks. Survey reports of hours worked per week correlated in the range of .60 to .64 with the corresponding records information. Reports of hourly wages correlated even less with company records—for the most recent pay period the correlation of the survey measure of hourly earnings with the records measure is .35, which means that only about 10 percent of the variance is valid variance.

We may assume that some difficulties of measurement are universal in the sense that they represent problems that must be surmounted in all surveys, i.e., peoples' memories are poor regardless of time and place. Although one would expect to be able to obtain a higher degree of accuracy in the measurement of such things as recent employment and earnings information, some phenomena are clearly not amenable to valid and reliable retrospective measurement. For example, Kessler, Mroczek, and Belli (1994) suggested in their assessment of retrospective measurement of childhood psychopathology that while it might be possible to obtain some long-term memories of salient aspects of childhood psychiatric disorders many childhood memories are lost due to either their lower salience or active processes of repression.

Another one of the assumptions listed in the foregoing that seems to be repeatedly violated in survey practice is that the respondent is willing to put forth the effort that is needed to provide accurate information to the researcher. There is a clear recognition in the survey methodology literature that an important component in generating maximally valid data is the fostering of a commitment on the part of the respondent to report information as accurately as possible (see Cannell, Miller, and Oksenberg, 1981). It is also clear from discussions of respondent burden that high cognitive and motivational demands placed on respondents may result in a reduction in the quality of data. The potential usefulness of these developments can be seen from Krosnick and Alwin's (1989) review of possible applications of the idea that respondents maximize their utilities [see Esser (1986); also see the discussion below]. One way of phrasing the issue relies on Herbert Simon's (1957; Simon and Stedry, 1968) concepts of *satisficing* and *optimizing* behavior. The term "satisficing" refers to expenditures of the minimum amount of effort necessary to generate a satisfactory response to a survey question, in contrast to expenditures of a great deal of effort to generate a maximally valid response, or "optimizing" (see Tourangeau, 1984; Krosnick and Alwin, 1988, 1989; Alwin, 1991). We suggest that satisficing seems to be the most likely to occur when the costs of optimizing are high, which is a function of three general factors: the inherent difficulty of the task, the respondent's capacities or abilities to perform the task, and the respondent's motivation to perform

the task. There is plenty of evidence in the survey methods literature that provides a basis for a "satisficing" interpretation of measurement errors: random responding (Converse, 1964, 1970, 1974), the effects of Don't Know filters (Schuman and Presser, 1981; McClendon and Alwin, 1993), the effects of offering middle alternatives (Schuman and Presser, 1981), response order effects (Krosnick and Alwin, 1987), acquiescence response bias (McClendon, 1991), and nondifferentiation in the use of rating scales (Krosnick and Alwin, 1988). The conditions under which this wide array of well-documented response errors are precisely those that are known to foster satisficing (Krosnick and Alwin, 1989; Krosnick, 1999).

To take one further example of areas in which one can expect measurement errors to occur, consider the reporting of socially undesirable or embarassing events. There has been considerable focus in the survey methods literature on underreporting problems with deviant and socially undesirable behavior (Bradburn, Rips, and Shevell, 1987; Cannell, Miller and Oksenberg, 1981). Also, on the flip side is the problem of socially desirable behaviors, e.g., church attendance in the United States or voting in national elections which tend to be over reported (see Hadaway, Marler, and Chaves, 1993; Traugott and Katosh, 1979, 1981; Presser, 1990; Presser and Traugott, 1992). Clearly what is socially desirable or undesirable in one segment of the population is not necessarily so in another, so a systematic understanding of this set of issues is necessary within the framework of a consideration of measurement errors.

All of these examples are instances in which there is a difference between the quantity or quality *recorded or observed* and the *true value* of the variable. Obviously, in order to address problems of measurement error it is important to understand what is meant by the concept of the "true" value, since error is defined in relation to it. Approaches to dealing with measurement error differ in how to conceptualize the "true" value (see Groves, 1989, 1991). To simplify matters, there are at least two different conceptions of error, based on two different conceptions of true score. "Platonic" true scores are those for which there is some notion of "truth" as is usually assumed in record-check studies (see Marquis and Marquis, 1977). Studies, for example, that compare voting records with self-reports of voting (Presser and Traugott, 1992) or those that compare actual patterns of religious behavior compared with self-reports (Hadaway et al., 1993) are implicitly using a *platonic* notion of true scores, i.e., that there is some "truth" out there that is to be discovered. This definition, however, is not generally useful because most variables we wish to measure in surveys have no "true" source against which to measure the accuracy of the survey report. "Psychometric" true scores, on the other hand, are defined in statistical terms as the expected value of a hypothetical infinite set of observations for a *fixed person* (see Lord and Novick, 1968). We return to a discussion of these matters in the following chapters dealing with the estimation of components of measurement error.

2.2 SOURCES OF MEASUREMENT ERROR IN SURVEY REPORTS

The claim that there are errors in survey reports suggests that something happens during the process of gathering data that creates this disparity between the "observed"

value and the "true" value. We can perhaps begin to understand the potential errors in survey measurement if we make explicit the assumptions that are often made in the collection of survey data. These are essentially as follows: (1) that the question asked is an appropriate and relevant one, which has an answer; (2) that the question is posed in such a way that the respondent or informant understands the information requested; (3) that the respondent or informant has access to the information requested; (4) that the respondent or informant can retrieve the information from memory; (5) that respondents are motivated to make the effort to provide an accurate account of the information retrieved; and (6) that they can communicate this information into the response framework provided by the survey question. Obviously, it would be naive to assume that these assumptions are met in every case, and so to acknowledge the possibility that they are not opens the door to the conclusion that measurement errors may occur in the gathering of survey data.

As noted earlier, there are *two* fundamental strategies to dealing with measurement errors in surveys. The first is to concentrate on the aspects of the *information-gathering process* that contribute to errors and reduce them through improved data collection methods. The second is to accept the fact that measurement errors are bound to occur, even after doing everything that is in ones power to minimize them, and *to model the behavior of errors using statistical designs.* We return to the latter topic in a subsequent chapter. Here we focus on ways in which survey measurement errors can be reduced by knowing about when and how they tend to occur.

Regardless of whether one's focus is on the reduction of errors during the collection of survey data or on the modeling of errors once they have occurred, there is a common interest between these emphases in developing an accurate picture of the response process and the factors that impinge on the collection of accurate information in the survey interview. Responses to survey questions are affected by a number of factors that produce the types of errors of measurement discussed above. It is generally agreed that key sources of measurement errors are linked to aspects of survey questions, the cognitive processes of information processing and retrieval, the motivational context of the setting that produces the information, and the response framework in which the information is then transmitted (see Alwin, 1989, 1991; Alwin and Krosnick, 1991b; Bradburn and Danis, 1984; Cannell, et al., 1981; Hippler, Schwarz and Sudman, 1987; Knäuper, Belli, Hill, and Herzog, 1997; Krosnick, 1999; Schwarz, 1999a, 1999b; Schwarz and Sudman, 1994; Sirken, Herrmann, Schechter, Schwarz, Tanur, and Tourangeau, 1999; Strack and Martin, 1987; Tourangeau, 1984, 1987, 1999; Tourangeau and Rasinski, 1988).

A classic treatment of this issue by Oksenberg and Cannell (1977), for example, examines the logical flow of steps by which the individual respondent processes the information requested by a question. Their work has influenced the following discussion, and for our purposes there are essentially *six* critical elements of the response process that directly impinge on the reliability and validity of survey measurement to which we devote attention in this section (see Table 2.1). All of these factors play a role in affecting the quality of survey data, whether the question seeks information of a factual nature, or whether it asks for reports of subjective states, such as beliefs and attitudes, but they are perhaps especially problematic in the

Table 2.1. Six critical elements in the response process

1.	*Question validity:* the adequacy of the question in measuring the phenomenon of interest
2.	*Comprehension:* the respondent's understanding or comprehension of the question and the information it requests
3.	*Accessibility:* the respondent's access to the information requested (e.g., do they have an opinion?)
4.	*Retrieval:* the respondent's capacities for developing a response on the basis of the information at hand, say, from internal cognitive and affective cues regarding his/her attitude or level of approval
5.	*Motivation:* the respondent's willingness to provide an accurate response
6.	*Communication:* the respondent's translation of that response into the response categories provided by the survey question

measurement of subjective phenomena such as attitudes, beliefs, and self-perceptions. It is particularly important to "deconstruct" the process of questioning respondents and recording their answers within a temporal framework such as this in order to be able to identify sources of survey error.

2.2.1 Validity of the Question

The concept of *measurement validity* in its most general sense refers to the extent to which the measurement accomplishes the purpose for which it is intended. This set of considerations, thus, expresses a concern with the linkage between concepts and their indicators. For example, according to the *Standards for educational and psychological testing* (APA, 2000), we can usefully distinguish among three types of validity in this sense—content, criterion-related, and construct validity. *Content validity* refers to the extent to which a well-specified conceptual domain has been represented by available or selected measures. Too often, survey researchers do not really know what they want to measure, and they therefore beg the question of *measurement validity* from the outset. This is often seen in the way they select questions for use in surveys, namely, going to other people's surveys to see what other people have asked. This is apparently done on the assumption that the other person "knows" what they are measuring. Nothing could probably be further from the truth. Another indication of the survey researcher's opting out of the concern with valid measurement is to include several questions on the same topic, apparently on the assumption that one or more of the questions will cohere around something important. This is why factor analysis is often used after a pretest to examine whether in fact that coherence has been achieved. Generally speaking, if one knows what one wants to measure, it will take no more than a few questions to get at it.