

## 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 peof! . . . 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

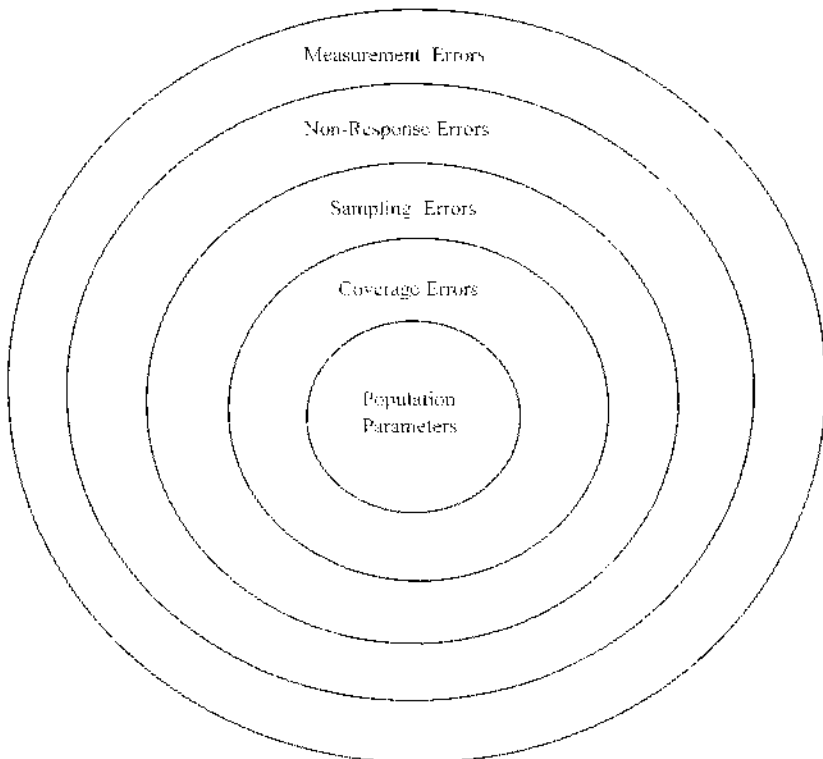


Figure 1.1. The relationship of sources of survey errors.

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 ( $\mu$ ), that is,  $MSL(\bar{y}) = E(\bar{y} - \mu)^2$ . This results in the standard expression:

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

where *Bias*<sup>2</sup> refers to the square of the theoretical quantity  $\bar{y} - \mu$ , and *Variance* refers to the variance of the sample mean  $\sigma_{\bar{y}}^2$ . 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 <sup>2</sup>	+	Variance
		⏟		⏟
		<i>Nonobservational Errors</i>		<i>Nonobservational Errors</i>
		Coverage bias		Coverage area variance
		Sampling bias		Sampling error variance
		Nonresponse bias		Nonresponse error variance
		<i>Observational Errors</i>		<i>Observational Errors</i>
		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).

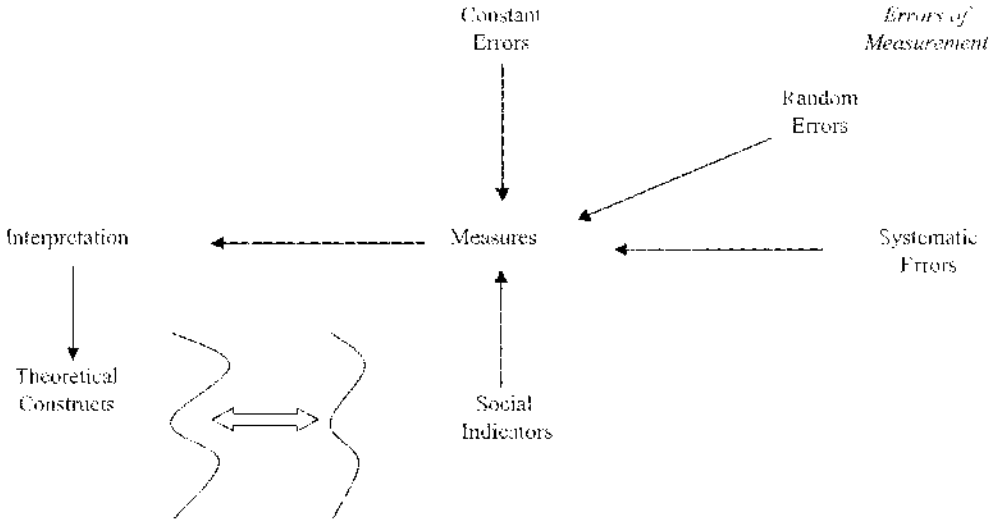


Figure 1.2. The relationship between constructs, indicators, measures, and measurement errors.

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 ( $\alpha$ )], 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  $\alpha$  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.