Chapter IX
Measurement error in household surveys: sources and measurement

Daniel Kasprzyk
Mathematica Policy Research
Washington, D.C., United States of America

Abstract

The present chapter describes the primary sources of measurement error found in sample surveys and the methods typically used to quantify measurement error. Four sources of measurement error - the questionnaire, the data-collection mode, the interviewer, and the respondent - are discussed, and a description of how measurement error occurs in sample surveys through these sources of error is provided. Methods used to quantify measurement error, such as randomized experiments, cognitive research studies, repeated measurement studies, and record check studies, are described and examples are given to illustrate the application of the method.

Key terms: measurement error, sources of measurement error, methods to quantify measurement error.

A. Introduction

1. Household survey data are collected through a variety of methods. Inherent in the process of collecting these data is the assumption that the characteristics and concepts being measured may be precisely defined, can be obtained through a set of well-defined procedures, and have true values independent of the survey. Measurement error is then the difference between the value of a characteristic provided by the respondent and the true (but unknown) value of that characteristic. As such, measurement error is related to the observation of the variable through the survey data-collection process, and, consequently, is sometimes referred to as an "observation error" (Groves, 1989).

2. The present chapter is based on a chapter on measurement error in a working paper prepared by a subcommittee on measuring and reporting the quality of survey data of the United States Federal Committee on Statistical Methodology (2001). As such, many of the references and examples refer to research in the United States of America and other developed countries. Nevertheless, the discussion applies to all surveys, no matter where they are conducted. The chapter should therefore be equally useful for those conducting surveys in developing and transition countries.

3. A substantial literature exists on measurement error in sample surveys [see Biemer and others (1991) and Lyberg and others (1997)] for reviews of important measurement error issues. Measurement error can give rise to both bias and variable errors (variance) in a survey estimate over repeated trials of the survey. Measurement bias or response bias occurs as a systematic pattern or direction in the difference between the respondents' answers to a question and the true values. For example, respondents may tend to forget to report income earned from a second or third job held, resulting in reported incomes lower than the actual incomes for some respondents. Variance occurs if values are reported differently when questions are asked more than once over the units (households, people, interviewers, and questionnaires) that are the sources of errors. Simple response variance reflects the random variation in a respondent's answer to a survey question over repeated questioning (that is to say, respondents may provide different answers to the same question if they are asked the question several times). The variable effects interviewers may have on the respondents' answers can be a source of variable error, termed interviewer variance. Interviewer variance is one form of correlated response variance that occurs because response errors are correlated for sample units interviewed by the same interviewer.

4. Several general approaches for studying measurement error are evident in the literature. One approach compares the survey responses with potentially more accurate data from another source. The data could be at the individual sample unit level as in a "record check study". As a simple example, if respondents were asked their ages, responses could be verified against birth records. However, we need to recognize that, even in this simple case, one cannot assume for certain that birth records are without errors. Nonetheless, one method of studying measurement error in a sample survey is to compare survey responses with data from other independent and valid sources. An alternative means of assessing measurement error using data from another source is to perform the analysis at the aggregate level, that is to say, to compare the survey-based estimates with population estimates from the other source. A second approach involves obtaining repeated measurements on some of the sample units. This typically is a survey

reinterview programme and involves comparing responses from an original interview with those obtained in a second interview conducted soon after the original interview. A third approach to studying measurement error entails selecting random subsamples of the full survey sample and administering different treatments, such as alternative questionnaires or questions or different modes of data collection. Finally measurement error can also be assessed in qualitative settings. Methods include focus groups and controlled laboratory settings, such as the cognitive research laboratory.

5. This chapter describes the primary sources of measurement error found in sample surveys and their measurement. Setting up procedures to quantify measurement error is expensive and often difficult to implement. For this reason and because it is good practice, survey managers place more emphasis on attempting to control the sources of measurement error though good planning and good survey implementation practices. Such practices include testing of survey materials, questionnaires and procedures, developing and testing well-defined, operationally feasible survey concepts, making special efforts to address data-collection issues for difficult-to-reach subgroups, implementing high standards for the recruitment of qualified field staff, and developing and implementing intensive training programmes and well-specified and clearly written instructions for the field staff. The control of non-sampling error, and measurement error specifically, requires an extended discussion by itself. See, for example, the report issued by the United Nations (1982) that includes a "checklist" for controlling non-sampling error in household surveys. This chapter does not address this issue, but rather focuses on describing the key sources of measurement error in sample surveys, and the typical ways measurement error is quantified.

6. Following Biemer and others (1991), four sources of error will be discussed: the questionnaire, the data-collection mode, the interviewer, and the respondent. A significant portion of the chapter describes how measurement error occurs in sample surveys through these sources of error. It then discusses some approaches to quantifying measurement error. These approaches include randomized experiments, cognitive research studies, repeated measurement studies, and record check studies. Quantifying measurement error always requires taking additional steps prior to, during, and after the conduct of survey. Frequently cited drawbacks to initiating studies that quantify specific sources of measurement error are the time and expense required to conduct the study. However, studies of measurement error are extremely valuable both to quantify the level of error in the current survey and to indicate where improvements should be sought for future surveys. Such studies are particularly useful for repeated survey programmes.

B. Sources of measurement error

7. Biemer and others (1991) identify four primary sources of measurement error:

- *Questionnaire*: the effect of the questionnaire design, its visual layout, the topics it covers, and the wording of the questions.

- *Data-collection method*: the effect of how the questionnaire is administered to the respondent (for example, mail, in person, or diary). Respondents may answer

questions differently in the presence of an interviewer, by themselves, or by using a diary.

- *Interviewer*: the effect that the interviewer has on the response to a question. The interviewer may introduce error in survey responses by not reading the items as intended, by probing inappropriately when handing an inadequate response, or by adding other information that may confuse or mislead the respondent.

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

8. These four sources are critical in the conduct of a sample survey. The questionnaire is the method of formally asking the respondent for information. The data-collection mode represents the manner in which the questionnaire is delivered or presented (self-administered or in person). The interviewer, in the case of the in-person mode, is the deliverer of the questionnaire. The respondent is the recipient of the request for information. Each can introduce error into the measurement process. Most surveys look at these sources separately, that is to say, if they address them at all. The sources can, however, interact with each other, for example, interviewers' and respondents' characteristics may interact to introduce errors not be evident from either source alone. The ways in which measurement error may arise in the context of these four error sources are discussed below.

1. Questionnaire effects

9. The questionnaire is the data collector's instrument for obtaining information from a survey respondent. During the last 20 years, the underlying principles of questionnaire design, once thought to be more art than science, have become the subject of an extensive literature (Sirken and others, 1999; Schwarz, 1997; Sudman, Bradburn, and Schwarz, 1996; Bradburn and Sudman, 1991). The questionnaire or the characteristics of the questionnaire, that is to say, the way the questions are worded or the way the questionnaire is formatted may affect how an individual responds to the survey. In the present section, we describe ways in which the questionnaire can introduce error into the data-collection process.

Specification problems

10. In the planning of a survey, problems often arise because research objectives and the concepts and information collected in the questionnaire are ambiguous, not well defined, or inconsistent. The questions in the questionnaire as formulated may be incapable of eliciting the information required to meet the research objectives. Data specification problems can arise because questionnaires and survey instructions are poorly worded, because definitions are ambiguous, or because the desired concept is difficult to measure. For example, a survey could ask about "the maternity care received during pregnancy" but not specify either which pregnancy or which period of time the question relates to. Ambiguity may arise in questions as basic as, how many jobs do you have?, if the nature of the job -- temporary or permanent jobs and/or full- or part-time -- is unspecified. Composite analytical concepts, such as total income for a person,

may not be reported completely if the individual components of income are not identified and defined for the respondent.

Question wording

11. The questions in the survey questionnaire must be precisely and clearly worded if the respondent is to interpret the question as the designer intended. Since the questionnaire is a form of communication between the data collector and the respondent, there are many potential sources of error. First, the questionnaire designer may not have clearly formulated the concept he/she is trying to measure. Next, even if the concept is clearly formulated, it may not be properly represented in the question or set of questions; and even if the concept is clear and faithfully represented, the respondent's interpretation may not be that intended by the questionnaire designer. Language and cultural differences or differences in experience and context between the questionnaire designer and the respondent may contribute to a misunderstanding of the questions. These differences can be particularly important in developing and transition countries that have several different ethnic groups. Vaessen and others (1987) discuss linguistic problems in conducting surveys in multilingual countries.

12. There are at least two levels in the understanding of a question posed in a sample survey. The first level is that of the simple understanding of the question's literal meaning. Is the respondent familiar with the words included in the question? Can the respondent recall information that matches his/her understanding of those words and provide a meaningful response? To respond to a question, however, the respondent must also infer the questionnaire's intent; that is to say, to answer the question, the respondent must determine the pragmatic meaning of the question (Schwarz, Groves and Schuman, 1995). It is this second element that makes the wording of questions a more difficult and more complex task than that of just constructing items requiring a low reading level. To produce a well-designed instrument, respondents' input, that is to say, their interpretation and understanding of questions, is needed. Cognitive research methods offer a useful means of obtaining this input (see sect. C.2).

Length of the questions

13. Common sense and good writing practice suggest that keeping questions short and simple will lead to clear interpretation. Research finds, however, that longer questions may elicit more accurate detail from respondents than shorter questions, at least in respect of reporting behaviour as related to symptoms and doctor visits (Marquis and Cannell, 1971) and alcohol and drug use (Bradburn, Sudman and Associates, 1979). Longer questions may provide more information or cues to help the respondent remember and more time to think about the information being requested.

Length of the questionnaire

14. Researchers and analysts always want to ask as many questions as possible, while the survey methodologist recognizes that error may be introduced if the questionnaire is too long. A respondent can lose concentration or become tired depending on his/her characteristics (age or

health status, for example), salience of the topic, rapport with the interviewer, design of the questionnaire, and mode of interview.

Order of questions

15. Researchers have observed that the order of the questions affects response (Schuman and Presser, 1981), particularly in attitude and opinion surveys. Assimilation -- where subsequent responses are oriented in the same direction as those for preceding items, and contrast, where subsequent responses are oriented in the opposite direction from those for preceding items -- has been observed. Respondents may also use information derived from previous items regarding the meaning of terms to help them answer subsequent items.

Response categories

16. Question response categories may affect responses by suggesting to the respondent what the developer of the questionnaire thinks is important. The respondent infers that the categories included with an item are considered to be the most important ones by the questionnaire developer. This can result in confusion as to the intent of the question if the response categories do not appear appropriate to the respondent. The order of the categories may also affect responses. Respondents may become complacent during an interview and systematically respond at the same point on a response scale, respond to earlier choices rather than later ones, or choose the later responses offered.

17. The effect produced by the order of the response categories may also be influenced by the mode in which the interview is conducted. If items are self-administered, response categories appearing earlier in the list are more likely to be recalled and agreed with (primacy effect), because there is more time for the respondent to process them. If items are interviewer-administered, the categories appearing later are more likely to be recalled (recency effect).

Open and closed formats

18. A question format in which respondents are offered a specified set of response options (closed format) may yield different responses than that in which respondents are not given such options (open format) (Bishop and others, 1988). A given response is less likely to be reported in an open format than when included as an option in a closed format (Bradburn, 1983). The closed format may remind respondents to include something they would not otherwise remember. Response options may indicate to respondents the level or type of responses considered appropriate [see, for example, Schwarz, Groves and Schuman (1995) and Schwarz and Hippler (1991)].

Questionnaire format

19. The actual “look” of a self-administered questionnaire, that is to say, the questionnaire format and layout, may help or hinder accurate response. The fact that respondents may become confused by a poorly formatted questionnaire design could result in a misunderstanding of skip patterns, or contribute to misinterpretation of questions and instructions. Jenkins and Dillman

(1997) provide principles for designing self-administered questionnaires for the population of the United States. Caution should be exercised in transferring these principles to another country without having considered the cultural and linguistic factors unique to that country.

2. Data-collection mode effects

20. Identifying the most appropriate mode of data collection entails a decision involving a variety of survey methods issues. Financial resources often play a significant role in the decision; however, the content of the questionnaire, the target population, the target response rates, the length of the data-collection period, and the expected measurement error are all important considerations in the process of deciding on the most appropriate data-collection mode. While advances in technology have led to increases in the use of the telephone as a means of data collection, the number of other modes of data-collection offer substantial variety of options in the conduct of a survey. Lyberg and Kasprzyk (1991) present an overview of different data-collection methods along with the sources of measurement error for these methods. A summary of this overview is presented below.

Face-to-face interviewing

21. Face-to-face interviewing is the main method of data collection in developing and transition countries. In most cases, an interviewer administers a structured questionnaire to respondents and fills in the respondent's answers on the paper questionnaire. The use of this paper and pencil personal interview (PAPI) method has had a long history. Recent advances in the production of lightweight laptop personal computers have resulted in face-to-face interviewing conducted via computer-assisted personal interviewing (CAPI). Interviewers visit the respondents' home and conduct interviews using laptop computers rather than paper questionnaires. See Couper and others (1998) for a discussion of issues related to CAPI. The most obvious advantage of the CAPI methodology relates to quality control and the reduction of response error. Interviewers enter responses into a computer file. The interview software ensures that questionnaire skip patterns are followed correctly and that responses are entered and edited for reasonableness at the time of interview; as a result, time and resources are saved at the data cleaning stage of the survey.

22. With face-to-face interviewing, complex interviews may be conducted, visual aids may be used to help the respondent answer the questions, and skillful, well-trained interviewers can build rapport and probe for more complete and accurate responses. However, the interviewers may influence respondents' answers to questions, thereby producing a bias in the survey estimates or an interviewer variance effect as discussed in section C.3. Interviewers can affect responses through a combination of personality and behavioural traits. A particular concern relates to socially undesirable traits or acts. Respondents may well be reluctant to report such traits or acts to an interviewer. DeMaio (1984) notes that the factor of social desirability seems to encompass two elements – the idea that some things are good and others bad, and the fact that respondents want to appear "good" and will answer questions to appear so.

23. Another possible source of measurement error connected with face-to-face interviewing in household surveys is the possible presence of other household members at the interview.

Members of the household may affect a respondent's answers, particularly when the questions are viewed as sensitive. For example, it may be difficult for a respondent to answer questions related to the use of illegal drugs truthfully when another household member is present. Even seemingly innocuous questions may be viewed as sensitive in the presence of another household member (for example, marital or fertility history-related questions asked in the presence of a spouse).

Self-completion surveys

24. The sources of measurement error in self-completion surveys questionnaires are different from those in face-to-face interviewing. Self-administered surveys have, obviously, no interviewer effects and involve less of a risk of "social desirability" effects. They also provide a means of asking questions on sensitive or threatening topics without embarrassing the respondent. Another benefit is that they can, if necessary, be administered simultaneously to more than one respondent in a household (Dillman, 1983). On the other hand, self-completion surveys may suffer from systematic bias if the target population consists of individuals with little or no education, or individuals who have difficulty reading and writing. This bias may be observed in responses to "open-ended" questions which can be less thorough and detailed than those responses obtained in surveys conducted by interviewers. This method of data collection may be less than ideal in countries with low literacy rates; however, even if the target population has a reasonably high education level, respondents may misread and misinterpret questions and instructions. Generally, item response rates are lower in self-completion surveys, but when the questions are answered, the data tend to be of higher quality. Self-completion surveys, perhaps more than other data-collection modes, benefit from good questionnaire design and formatting and clearly written questionnaire items. One specific type of self-completion survey is the self-completion mail survey in which respondents are asked to complete by themselves a questionnaire whose delivery and retrieval is done by mail (Dillman, 1978; 1991; 2000).

Diary surveys

25. Diary surveys are self-administered forms used for topics that require detailed reporting of behaviour over a period of time (for example, e.g., expenditures, time use, and television viewing). To minimize or avoid recall errors, the respondent is encouraged to use the diary and record responses about an event or topic soon after its occurrence. The diary mode's success depends on the respondent's taking an active role in recording information and completing a typically "burdensome" form. This mode also entails the requirement that the target population be capable of reading and interpreting the diary questions, a condition that will not apply in countries with low literacy rates. The data-collection procedure usually requires that interviewers contact the respondent to deliver the diary, gain the respondent's cooperation and explain the data recording procedures. The interviewer returns after a predetermined amount of time to collect the diary and, if it has not been completed, to assist the respondent in completing it.

26. Lyberg and Kasprzyk (1991) identify a number of sources of measurement error for this mode. For example, respondents who pay little or no attention to recording events may fail to record events when fresh in their memories. The diary itself, because of its layout and format

and the complexity of the question items, may present the respondent with significant practical difficulties. Furthermore, respondents may change their behaviour as a result of using a diary; for example, the act of having to list purchases in an expenditure diary may cause a respondent to change his/her purchasing behaviour. Discussions of measurement errors in expenditure surveys and, in particular, the diary aspect of the surveys, can be found in Neter (1970) and Kantorowitz (1992). Comparisons of data derived from face-to-face interviews and diary surveys are found in Silberstein and Scott (1991).

Direct observation

27. Direct observation, as a data-collection method, requires the interviewer to collect data using his/her senses (vision, hearing, touching, testing) or physical measurement devices. This method is used in many disciplines, for example, in agricultural surveys to estimate crop yields ("eye estimation") and in household surveys to assess the quality of respondents' housing. Observers introduce measurement errors in ways similar to those through which errors are introduced by interviewers; for example, observers may misunderstand concepts and misperceive the information to be recorded, and may change their pattern of recording information over time because of complacency or fatigue.

3. Interviewer effects

28. The interviewer plays a critical role in many sample surveys. As a fundamental part of the data-collection process, his/her performance can influence the quality of the survey data. The interviewer, however, is one component of the collection process whose performance the survey researcher/survey manager can attempt to control; consequently, strategies have evolved--through selection and hiring, training, and monitoring of job performance -- to minimize the error associated with the role of the interviewer (Fowler, 1991). Because of individual differences, each interviewer will handle the survey situation in a different way; individual interviewers, for example, may not ask questions exactly as worded, follow skip patterns correctly or probe for answers in an appropriate manner. They may not follow directions exactly, either purposefully or because those directions have not been made clear. Without being aware, interviewers may vary their inflection or tone of voice, or display other changes in personal mannerisms.

29. Errors, both overreports and underreports, can be introduced by each interviewer. When overreporting and underreporting approximately cancel out across all interviewers, small overall interviewer bias will result. However, errors of individual interviewers may be large and in the same direction, resulting in large biases for those interviewers. Variation in the individual interviewer biases gives rise to what is termed interviewer variance, which can have a serious impact on the precision of the survey estimates.

Correlated interviewer variance

30. In the early 1960s, Kish (1962) developed an approach using the intra-interviewer correlation coefficient, which he denoted by ρ, to assess the effect of interviewer variance on survey estimates. The quantity ρ, which is defined as the ratio of the interviewer variance

component to the total variance of a survey variable, is estimated by a simple analysis of variance.

31. In well-conducted face-to-face surveys, ρ typically is about 0.02 for most variables. Although ρ is small, the effect on the precision of the estimate may be large. The variance of the sample mean is multiplied by $1 + \rho\,(n\text{-}1)$, where n is the size of the average interviewer workload. A ρ of 0.02 with a workload of 10 interviews increases the variance by 18 per cent, and a workload of 25 yields a variance 48 per cent larger. Thus, even small values of ρ can significantly reduce the precision of survey statistics. Based on practical and economic considerations, interviewers usually have large workloads. Thus, an interviewer who contributes a systematic bias will affect the results obtained from a sizeable number of respondents and the effect on the variance can be large.

Interviewer characteristics

32. The research literature is not helpful in identifying characteristics indicative of good interviewers. In the United Kingdom of Great Britain and Northern Ireland, Collins (1980) found no basis for recommending that the recruitment of interviewers should be concentrated among women rather than men, or among middle-class persons, or among the middle-aged rather than the young or the old. Weiss (1968), studying a sample of welfare mothers in New York City, validated the accuracy of several items, and found that the similarity between interviewer and respondent with respect to age, education and socio-economic status did not result in better reporting. Sudman and others (1977) studied interviewer expectations of the difficulty of obtaining sensitive information and observed weak effects in respect of the relationship between expected and actual interviewing difficulties. Groves (1989) reviewed a number of studies and concluded, in general, that demographic effects may occur when measurements are related to the demographic characteristics, but not otherwise; for example, there may be an effect based on the race of the interviewer if the questions are related to race.

Methods to control interviewer errors

33. To some extent, the survey manager can control interviewer errors through interviewer training, supervision or monitoring, and workload manipulation. A training programme of sufficient length to cover interview skills and techniques as well as provide information on the specific survey helps to bring a measure of standardization to the interview process (Fowler, 1991). Many believe standardizing interview procedures reduces interviewer effects.

34. Supervision and performance monitoring, the objectives of which are to monitor performance through observation and performance statistics and identify problem questions, constitute another component of an interviewer quality control system. Reinterview programmes and field observations are conducted to evaluate individual interviewer performance. Field observations are conducted using extensive coding lists or detailed observers' guides where the supervisor checks whether the procedures are properly followed. For instance, the observation could include the interviewer's appearance and conduct, the introduction of himself/herself and of the survey, the manner in which the questions are asked and answers recorded, the use of

show cards and neutral probes, and the proper use of the interviewers' manual. In other instances, tapes (either audio-visual or audio) can be made and interviewer behavior coded and analysed (Lyberg and Kasprzyk, 1991).

35. Another way to reduce the effect of interviewer variance is to lower the average workload; however, this assumes that additional interviewers of the same quality are available. Groves and Magilavy (1986) discuss optimal interviewer workload as a function of interviewer hiring and training costs, interview costs, and size of intra-interviewer correlation. Since the intra-interviewer correlation varies among statistics in the same survey, it is very difficult to ascertain what constitutes an optimal workload.

36. Interviewer effects can be reduced by avoiding questionnaire design problems, by giving clear and unambiguous instructions and definitions, by training interviewers to follow the instructions, and by minimizing reliance on the variable skills of interviewers with respect to obtaining responses.

4. Respondent effects

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

- *Encoding of information*, which involves the process of forming memories or retaining knowledge.

- *Comprehension of the survey question*, which involves knowledge of the questionnaire's words and phrases as well as the respondent's impression of the survey's purpose, the context and form of the question, and the interviewer's behaviour when asking the question.

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

- *Judgement of appropriate answer*, which involves the respondent's choice of alternative responses to a question based on the information that was retrieved;

- *Communication of the response*, which involves influences on accurate reporting after the respondent retrieved the relevant information and the respondent's ability to articulate the response.

38. Many aspects of the survey process affect the quality of the respondent's answers emerging from this five-stage process. Examples of factors that influence respondent effects follow.

Respondent rules

39. Respondent rules that define the eligibility criteria used for identifying the person(s) to answer the questionnaire play an important role in the response process. If a survey collects information about households, knowledge of the answers to the questions may vary among the different eligible respondents in the household. Surveys that collect information about individuals within sampled households may use self-reporting or proxy reporting. Self-reporting versus proxy reporting differences vary by subject matter (for example, self-reporting is better for attitudinal surveys). United Nations (1982) describes the result of a pilot test of the effects of proxy response on demographic items for the Turkish Demographic Survey. Blair, Menon, and Bickart (1991) present a literature review of research on self-reporting versus proxy reporting.

Questions

40. The wording and complexity of the question and the design of the questionnaire may influence how and whether the respondent understands the question (see sect. B.1 for further details). The respondent's willingness to provide correct answers is affected by the types of question asked, by the difficulty of the task in determining the answers, and by the respondent's view of the social desirability of the responses.

Interviewers

41. The interviewer's visual clues (for example, age, gender, dress, facial expressions) as well as audio cues (for example, tone of voice, pace, inflection) may affect the respondent's comprehension of the question.

Recall period

42. Time generally reduces ability to recall facts or events. Memory fades, resulting in respondents' having more difficulty recalling an activity when there is a long time period intervening between an event and the survey. For example, for some countries in the World Fertility Survey, recent births are likely to be dated more accurately than births further back in time (Singh, 1987). Survey designers may seek recall periods that minimize the total mean squared error in terms of the sampling error and possible biases; for example, Huang (1993) found the increase in precision obtained by increasing sample size and changing from a four-month reference period to a six-month reference period would not compensate for the increase in bias from recall loss. Eisenhower, Mathiowetz and Morganstein (1991) discuss the use of memory aids (for example, calendars, maps, diaries) to reduce recall bias. Mathiowetz (2000) reports the results of a meta-analysis testing the hypothesis that the quality of retrospective reports is a function of the length of recall period.

Telescoping

43. Telescoping occurs when respondents report an event as being within the reference period when it actually occurred outside that period. Bounding techniques (for example, conduct of an initial interview solely to establish a reference date, or use of a significant date or event as

the beginning of the reference period) can be used to reduce the effects of telescoping (Neter and Waksberg, 1964).

Panel/longitudinal surveys

44. Additional respondent-related factors contribute to survey error in panel or longitudinal surveys. First, spurious measures of change may occur when a respondent reports different answers to the same or similar questions at two different points and the responses are due to random variation in answering the same questions rather than real change. Kalton, McMillen and Kasprzyk (1986) provide examples of measurement error in successive waves of a longitudinal survey. They cite age, race, sex, and industry and occupation, as variables where measurement error was observed in the United States Survey of Income and Program Participation. The United States Survey of Income and Program Participation Quality Profile discusses this and other measurement error issues identified in the survey (United States Bureau of the Census, 1998). Dependent interviewing techniques, in which the responses from the previous interview are used in the current interview, can reduce the incidence of spurious changes. Hill (1994) found dependent interviewing had resulted in a net improvement in measures of change in occupation and industry of employment, but it can also miss reports of true change, so selectivity in its use is necessary. Mathiowetz and McGonagle (2000) review current practices within a computer-assisted interviewing environment as well as empirical evidence of the impact of dependent interviewing on data quality.

45. Panel conditioning or "time-in-sample" bias is another potential source of error in panel surveys. Conditioning refers to the change in response occurring when a respondent has had one or more prior interviews. Woltman and Bushery (1977) investigated time-in-sample bias for the United States National Crime Victimization Survey, comparing victimization reports of individuals with varying degrees of panel experience (that is to say, number of previous interviews) who had been interviewed in the same month. They found generally declining rates of reported victimization as the number of previous interviews increased. Kalton, Kasprzyk and McMillen (1989) also discuss this source of error.

C. Approaches to quantifying measurement error

46. There exist several general approaches to quantifying measurement error. In order to study measurement biases, different treatments, such as alternative questionnaires or questions or a different mode of data collection, can be administered to randomly selected subsamples of the full survey sample. Measurement error can be studied in qualitative settings, such as focus groups, or cognitive research laboratories. Another approach involves repeated measurements on the sample unit, such as are undertaken in a survey reinterview programme. Finally, there are "record check studies", which compare survey responses with more accurate data from another source to estimate measurement error. These approaches are discussed below.

1. Randomized experiments

47. A randomized experiment is a frequently used method for estimating measurement errors. Survey researchers have referred to this method by a variety of names such as interpenetrated samples, split-sample experiments, split-panel experiments, random half-sample experiments, and split-ballot experiments. Different treatments related to the specific error being measured are administered to random subsamples of identical design. For studying variable errors, many different entities thought to be the source of the error are included and compared (for example, many different interviewers for interviewer variance estimates). For studying biases, usually only two or three treatments are compared (for example, two different data-collection modes), with one of the methods being the preferred method. Field tests, conducted prior to conducting the survey, often include randomized experiments to evaluate alternative methods, procedures and questionnaires.

48. For example, a randomized experiment can be used to test the effect of the length of the questionnaire. Sample units are randomly assigned to one of two groups, one group receiving a "short" version of the questions and the other group receiving the "long" version. Assuming an independent data source is available, responses for each group can then be compared with the estimates from the data source, which is assumed to be accurate and reliable. Similarly, question order effects can be assessed by reversing the order of the question set in an alternate questionnaire administered to random samples. The method was used for a survey in the Dominican Republic, conducted as part of the worldwide Demographic and Health Surveys programme; the core questionnaire was used for two-thirds of the sample and the experimental questionnaire was used for one third of the sample. The goal was to determine response differences resulting from the administration of two sets of questions (Westoff, Goldman and Moreno, 1990).

2. Cognitive research methods

49. During the last 20 years, the use of cognitive research methods for the reduction of measurement error has grown rapidly. These methods were initially used to obtain insight into respondents' thought processes, but are increasingly used to supplement traditional field tests (Schwarz and Sudman, 1996; Sudman, Bradburn and Schwarz, 1996). Respondents provide information to the questionnaire designer on how they interpret the items in the questionnaire. This approach is labour-intensive and costly per respondent; consequently, cognitive testing is conducted on small samples. One weakness of cognitive interviews is that they are conducted with small non-random samples. The questionnaire designer must recognize that the findings reveal potential problems but are not necessarily representative of the potential survey respondents.

50. Most widely used methods rely on verbal protocols (Willis, Royston and Bercini, 1991). Respondents are asked to complete the draft questionnaire and to describe how they interpret each item. An interviewer will probe regarding particular words, definitions, skip patterns, or other elements of the questionnaire on which he or she wishes to obtain specific feedback from the respondent. Respondents are asked to identify anything not clear to them. Respondents may be asked to do this as they are completing the questionnaire ("concurrent think-aloud") or in a debriefing session afterwards ("retrospective think-aloud"). The designer may add probes to

investigate the clarity of different items or elements of the questionnaire in subsequent interviews. The advantage of the technique is that it is not subject to interviewer-imposed bias. The disadvantage is that it does not work well for respondents uncomfortable with, or not used to, verbalizing their thoughts (Willis, 1994).

51. A related technique involves the interviewer's asking the respondent about some feature of the question immediately after the respondent completes an item (Nolin and Chandler, 1996). This approach is less dependent on the respondent's comfort and skill level with respect to verbalizing his/her thoughts, but limits the investigation to those items the survey designer thinks he can ask about. The approach may also introduce an interviewer bias since the probes depend on the interviewer. Inasmuch as the probing approach is different from conducting an interview, some consider it artificial (Willis, 1994).

52. Other approaches allow the respondent to complete the survey instrument with questioning conducted in focus groups. Focus groups provide the advantage of the interaction of group members which may lead to the exploration of areas that might not be touched on in one-on-one interviews.

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

54. Cognitive research methods are now widely used in designing questionnaires and reducing measurement error in surveys in developed countries. Sudman, Bradburn and Schwarz (1996) summarize major findings as they relate to survey methodology. Tucker (1997) discusses methodological issues in the application of cognitive psychology to survey research.

3. Reinterview studies

55. A reinterview - a repeated measurement on the same unit in an interview survey - is an interview that asks the original interview questions (or a subset of them). Reinterviews are usually conducted with a small subsample (usually about 5 per cent) of a survey's sample units. Reinterviews are conducted for one or more of the following purposes:

- To identify interviewers who falsify data
- To identify interviewers who misunderstand procedures and require remedial training
- To estimate simple response variance
- To estimate response bias

56. The first two purposes provide information on measurement errors resulting from interviewer effects. The last two provide information on measurement errors resulting from the

joint effect of all four sources (namely, interviewer, questionnaire, respondent, and data-collection mode).

57. Specific design requirements for each of four types of reinterviews are discussed below [see Forsman and Schreiner (1991)]. In addition, some methods for analysing reinterview data along with limitations of the results are also presented.

Interviewer falsification reinterview

58. Interviewers may falsify survey results in several ways; for example, an interviewer can make up answers for some or all of the questions, or an interviewer can deliberately not follow survey procedures. To detect the occurrence of falsification, a reinterview sample is drawn and the reinterviews are generally conducted by supervisory staff. A falsification rate, defined as the proportion of interviewers falsifying interviews detected through the falsification reinterview, can be calculated. Schreiner, Pennie and Newbrough (1988) report a 0.4 per cent rate for the United States Current Population Survey, a 0.4 per cent rate for the United States National Crime Victimization Survey, and a 6.5 per cent rate for the New York City Housing and Vacancy Survey, which are all conducted by the United States Bureau of the Census.

Interviewer evaluation reinterview

59. Reinterview programmes that identify interviewers who do not perform at acceptable levels are called interviewer evaluation reinterviews. The purpose is to identify interviewers who misunderstand survey procedures and to target them for additional training. Most design features of this type of reinterview are identical to those of a falsification reinterview. Tolerance tables, based on statistical quality control theory, may be used to determine whether the number of differences in the reinterview after reconciliation exceeds a specific acceptable limit. Reinterview programmes at the United States Bureau of the Census use acceptable quality tolerance levels ranging between 6 and 10 per cent (Forsman and Schreiner, 1991).

Simple response variance reinterview

60. The simple response variance reinterview is an independent replication of the original interview procedures. All guidelines, procedures and processes of the original interview are repeated in the reinterview to the fullest extent possible. The reinterview sample is a representative subsample of the original sample design. The interviewers, data-collection mode, respondent rules and questionnaires of the original interview are used in the reinterview. In practice, the assumptions are not always followed; for example, if the original questionnaire is too long, a subset of the original interview questionnaire is used. Differences between the original interview and the reinterview are *not* reconciled.

61. A statistic estimated from a simple response variance reinterview is the gross difference rate (*GDR*), which is the average squared difference between the original interview and reinterview responses. The *GDR* divided by 2 is an unbiased estimate of simple response variance (*SRV*). For characteristics that have two possible outcomes, the *GDR* is equal to the percentage of cases that had different responses in the original interview and the reinterview.

Brick, Rizzo and Wernimont (1997) provide general rules for interpreting the response variance measured by the *GDR*.

62. Another statistic is the index of inconsistency (*IOI*), which measures the proportion of the total population variance attributed to the simple response variance. Hence,

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

where s^2_1 is the sample variance for the original interview and s^2_2 is the sample variance for the reinterview.

63. The value of the *IOI* is often interpreted as follows:

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

64. The response variance measures, the *GDR* and the *IOI*, provide data users with information on the reliability and response consistency of a survey's questions. Examples of the use of the *GDR* and the *IOI* for selected variables from a fertility survey in Peru can be found in United Nations (1982) on non-sampling error in household surveys. As part of the second phase of the Demographic and Health Surveys programme, a reinterview programme to assess the consistency of responses at the national level was conducted in Pakistan on a subsample of women interviewed in the main survey (Curtis and Arnold, 1994). Westoff, Goldman and Moreno (1990) describe a reinterview study conducted as part of the Demographic and Health Surveys programme in the Dominican Republic, notable because of the need to adopt several compromises, such as restricting the reinterviews to a few geographical areas and a subset of the target population. Reinterview surveys in India, conducted with a response variance objective, are described in United States Bureau of the Census (1985), which examines census evaluation procedures.

65. Feindt, Schreiner and Bushery (1997) describe a periodic survey's efforts to continuously improve questionnaires using a reinterview programme. When questions have high discrepancy rates as identified in the reinterview, questionnaire improvement research using cognitive research methods can be initiated. These methods may identify the cause of the problems and suggest possible solutions. During the next round of survey interviews, a reinterview can be conducted on the revised questions to determine whether reliability improvements have been made. This process is then repeated for the remaining problematic questions.

Response bias reinterview

66. A reinterview to measure response bias aims to obtain the true or correct responses for a representative subsample of the original sample design. In order to obtain the true answers, the

most experienced interviewers and supervisors are used. In addition, either the reinterview respondent used is the most knowledgeable respondent or the household members answer questions for themselves. The original interview questions are used for the reinterview, and the differences between the two responses are reconciled with the respondent to establish "truth." Another approach uses a series of probing questions to replace the original questions in an effort to obtain accurate responses and then reconcile differences with the respondent. For a discussion of reinterview surveys conducted with the objective of obtaining estimates of response bias, see the report describing census evaluation procedures issued by the United States Bureau of the Census (1985).

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

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

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

where n is the reinterview sample size; y_o is the original interview response; and y_T is the reinterview response after reconciliation, assumed to be the true response.

69. The *NDR* provides information about the accuracy of a survey question and also identifies questions providing biased results. The existence of this bias needs to be considered when the data are analysed and results interpreted. Brick and others (1996) used an intensive reinterview to obtain a better understanding of the respondent's perspective and reasons for his/her answers, leading to estimates of response bias. Although working with a small sample, the authors concluded that the method had potential for detecting and measuring biases. Bias-corrected estimates were developed, illustrating the potential effects on estimates when measures of bias are available.

4. Record check studies

70. A record check study compares survey responses for individual sample cases with values obtained from an external source, generally assumed to contain the true values for the survey variables. Such studies are used to estimate response bias resulting from the combined effect of all four sources of measurement error (interviewer, questionnaire, respondent and data-collection mode).

71. Groves (1989) describes the three kinds of record check study designs:

- The reverse record check
- The forward record check
- The full design record check

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

73. Often the record source is a listing of units (households or persons) with a given characteristic, such as those receiving a particular form of government aid. In this case, a reverse record check study does not measure overreporting errors (that is to say, units reporting the characteristic when they do not have it). These studies can measure only the proportion of the sample source records that correctly report or incorrectly do not report the characteristic. For example, a reverse record check study was conducted by the United States Law Enforcement Assistance Administration (1972) to assess errors in reported victimization. Police department records were sampled and the victim on the record was contacted. During the survey interview, the victims reported 74 per cent of the known crimes from police department records.

74. In a forward record check study, external record systems containing accurate information on the survey respondents are searched after the survey responses have been obtained. Response bias estimates are based on a comparison of survey responses with the values in the record systems. Forward record check studies provide the opportunity to measure overreporting. One difficulty with these kinds of studies is that they require contacting record-keeping agencies and obtaining permission from the respondents to obtain this information. If the survey response indicates that the unit does not have a given characteristic, it may be difficult to search the record system for that unit. Thus forward record check studies are limited in their ability to measure underreporting. Chaney (1994) describes a forward record check study for comparing teachers' self-reports of their academic qualifications with college transcripts. The data indicated that self-reports of types and years of degrees earned and major field were, for the most part, accurate; however, the reporting of courses and credit hours was less accurate.

75. A full design record check study combines features of both the reverse and forward record check designs. A sample is selected from a frame covering the entire population and records from all sources relevant to the sample cases are located. As a result, errors associated with underreporting and overreporting can be measured by comparing survey responses with all records (that is to say, from the sample frame as well as from external sources) for the survey respondents. Although this type of record check study avoids the weakness of the reverse and forward record check studies, it does require a database that covers all units in the population and all the corresponding events for those units. Marquis and Moore (1990) provide a detailed description of the design and analysis of a full record check study conducted to estimate measurement errors in the United States Survey of Income and Program Participation. In this study, survey data on the receipt of programme benefit amounts for eight Federal and State benefit programmes in four States were matched against the administrative records for the same

programmes. The Survey Quality Profile (United States Bureau of the Census, 1998) provides a summary of the design and analysis.

76. The three types of record check studies share limitations linked to the following three assumptions that, in practice, are unrealistic and are never justified: first, that record systems are free of errors of coverage, non-response, or missing data; second, that individual records in these systems are complete, accurate and free of measurement errors; and third, that matching errors (errors that occur as part of the process of matching the respondents' survey records) are non-existent or minimal.

Response bias for a given characteristic can be estimated by the average difference between the survey response and the record check value for that characteristic, according to the following formula:

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

where: n is the record check study sample size; Yi = the survey response for the i^{th} sample person; and X_i = the record check value for the i^{th} sample person.

78. The response bias measures from a record check study provide information about the accuracy of a survey question and identify questions that produce biased estimates. These measures can also be used for evaluating alternatives for various survey design features such as questionnaire design, recall periods, data-collection modes, and bounding techniques. For example, Cash and Moss (1972) give the results of a reverse record check study in three counties of North Carolina regarding motor vehicle accident reporting. Interviews were conducted in households containing sample persons identified as involved in motor vehicle accidents in the 12-month period prior to the interview. The study showed that whereas only 3.4 per cent of the accidents occurring within 3 months of the interview had not been reported, over 27 per cent of those occurring between 9 and 12 months before the interview had not been reported.

5. Interviewer variance studies

79. To study interviewer variance, interviewer assignments must be randomized so that differences in results obtained by different interviewers can be attributed to the effects of the interviewers themselves.

80. Interviewer variance is estimated by assigning each interviewer to different but similar respondents, that is to say, respondents who have the same attributes with respect to the survey variables. In practice, this equivalency is assured through randomization. The sample is divided into random subsets, each representing the same population, and each interviewer then works on a different subset of the sample. With this design, each interviewer conducts a small survey with all the essential attributes of the large survey except its size. O'Muircheartaigh (1982) describes the methodology used in the World Fertility Survey to measure the response variance due to interviewers and provides estimates of the response variance for the surveys conducted in Peru (1984a) and Lesotho (1984b).

81. In face-to-face interview designs, interpenetrated interviewer assignments are geographically defined to avoid large travelling costs. The assigned areas have sizes sufficient for one interviewer's workload. Pairs of assignment areas are identified and assigned to pairs of interviewers. Within each assignment area, each interviewer of the pair is assigned a random half of the sample housing units. Thus, each interviewer completes interviews in two assignment areas and each assignment area is handled by two different interviewers. The design consists of one experiment (a comparison of results of two interviewers in each of two assignment areas) replicated as many times as there are pairs of interviewers. Bailey, Moore, and Bailar (1978) present an example of interpenetration for personal interviews in the United States National Crime Victimization Survey in eight cities.

6. Behaviour coding

82. Interviewer performance, while both in training and on-the-job, can be evaluated through the use of behaviour coding. Trained observers observe a sample of interviews, code aspects of the interviews or the sample of interviews are tape-recorded and the coding is done from the tapes. Codes are assigned to record interviewer's major verbal activities and behaviours such as question asking, probe usage, and response summarization. For example, codes can classify how the interviewer reads the question, whether questions are asked correctly and completely, whether the questions are asked with minor changes and omissions, and whether the interviewer rewords the question substantially or does not complete the question. The coding system classifies whether probes directed the respondent to a particular response, further defined the question or were non-directive, whether responses were summarized accurately or inaccurately, and whether various other behaviours were appropriate or inappropriate. The coded results reflect to what extent the interviewer employed methods in which he/she had been trained, that is to say, an "incorrect" or "inappropriate" behaviour is defined as one that the interviewer had been trained to avoid. To establish and maintain a high level of coding reliability for each coded interview, a second coder should independently code a subsample of interviews.

83. A behaviour coding system can tell new interviewers which of their interviewing techniques are acceptable and which are not and may serve as a basis upon which interviewers and supervisors can review fieldwork and discuss the problems identified by the coding. Furthermore, it provides an assessment of an interviewer's performance, which can be compared both with the performance of other interviewers and with the individual's own performance during other coded interviews (Cannell, Lawson, and Hauser, 1975).

84. Oksenberg, Cannell and Blixt (1996) describe a study in which interviewer behaviour was tape-recorded, coded, and analysed for the purpose of identifying interviewer and respondent problems in the 1987 National Medical Expenditure Survey conducted by the United States Agency for Health Care Research and Quality. The study intended to see whether interview behaviour had differed from the principles and techniques covered in the interviewers' training. The authors reported that interviewers frequently had not asked the questions as worded, and at times they had asked them in ways that could influence responses. Interviewers had not probed as much as necessary; and when they did, the probes tended to be directive or inappropriate.

D. Concluding remarks: measurement error

85. Measurement error occurs through the data-collection process. Four primary sources were identified as being part of that process: the questionnaire, the method or mode of data collection, the interviewer, and the respondent. Quantifying the existence and magnitude of a specific type of measurement error requires advance planning and thoughtful consideration. Unless small-scale (that is to say, limited sample) studies are conducted, special studies are necessary that require randomization of subsamples, reinterviews, and record checks. These studies are usually expensive to conduct and require a statistician for the data analysis. Nevertheless, if there is sufficient concern that the issue may not be adequately resolved during survey preparations or if the source of error is particularly egregious in the survey being conducted, survey managers should takes steps to design special studies to quantify the principal or problematic source of error.

86. The importance of conducting studies to understand and quantify measurement error in a survey cannot be overemphasized. This is particularly critical if the survey concepts being measured are new and complicated. The analyses that users conduct are dependent on their having both good-quality data and an understanding of the nature and limitations of the data. Measurement error studies require an explicit commitment of the survey programme, because they are costly and time-consuming. The commitment, however, does not end with the implementation and conduct of the studies. The studies must be analysed and results reported so that analysts can make their own assessment of the effect of measurement error on their results. Special studies that focus on analyses of tests and experiments and assessments of data quality are typically available in methodological and technical reports [see, for example, methodological and analytical reports produced by the Demographic and Health Surveys program (Stanton, Abderrahim and Hill, 1997; Institute for Resource Development/Macro Systems Inc., 1990; Curtis, 1995)]. Finally, results from measurement error studies are important for improving the next fielding of the survey. Significant measurement improvements rely, to a large extent, on knowledge and results of previous surveys. Future improvements in the quality of survey data require the commitment of survey research professionals.

References

Bailey, L., T. F. Moore and B.A. Bailar (1978). An interviewer variance study for the eight impact cities of the National Crime Survey Cities Sample. *Journal of the American Statistical Association,* vol. 73, pp. 16–23.

Biemer, P.P., and G. Forsman (1992). On the quality of reinterview data with application to the current population survey. *Journal of the American Statistical Association*, vol. 87: pp. 915–923.

Biemer, P.P., and others , eds. (1991). *Measurement Errors in Surveys*. New York: John Wiley and Sons.

Bishop, G.F. and others (1988). A comparison of response effects in self-administered and telephone surveys. In *Telephone Survey Methodology*, R.M. Groves and others, eds. New York: John Wiley and Sons, pp. 321–340.

Blair, J., G. Menon and B. Bickart (1991). Measurement effects in self vs. proxy responses to survey questions: an information-processing perspective. In *Measurement Errors in Surveys,* P. Biemer and others, eds. New York: John Wiley and Sons, pp. 145–166.

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

_____ , and S. Sudman (1991). The current status of questionnaire design. In *Measurement Errors in Surveys*, P. Biemer and others, eds. New York: John Wiley and Sons, pp. 29-40.

_____ and Associates (1979). *Improving Interviewing Methods and Questionnaire Design: Response Effects to Threatening Questions in Survey Research*. San Francisco, California: Jossey-Bass.

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

Brick, J.M., and others (1996). *Estimation of Response Bias in the NHES: 95 Adult Education Survey*. Working Paper, No. 96-13. Washington, D.C., United States Department of Education, National Center for Education Statistics.

Cannell, C.F., S.A. Lawson and D.L. Hauser (1975). *A Technique for Evaluating Interviewer Performance*. Ann Arbor, Michigan: University of Michigan, Survey Research Center.

Cash, W.S., and A.J. Moss (1972). Optimum recall period for reporting persons injured in motor vehicle accidents. *Vital and Health Statistics*, vol. 2, No. 50. Washington, D.C.: Public Health Service.

Chaney, B. (1994). *The Accuracy of Teachers' Self-reports on Their Post Secondary Education: Teacher Transcript Study, Schools and Staffing Survey*. Working Paper, No. 94-04. Washington, D.C.: United States Department of Education, National Center for Education Statistics.

Collins, M. (1980). Interviewer variability: a review of the problem. *Journal of the Market Research Society*, vol. 22, No. 2, pp. 77–95.

Couper, M.P., and others, eds. (1998). *Computer Assisted Survey Information Collection*. New York: John Wiley and Sons.

Curtis, S.L. (1995). *Assessment of the Data Quality of Data Used for Direct Estimation of Infant and Child Mortality in DHS-II Surveys*. Occasional Papers, No. 3. Calverton, Maryland: Macro International, Inc.

_____ , and F. Arnold (1994). *An Evaluation of the Pakistan DHS Survey Based on the Reinterview Survey*. Occasional Papers, No. 1. Calverton, Maryland: Macro International, Inc.

Czaja R., and J. Blair (1996). *Designing Surveys: A Guide to Decisions and Procedures*. Thousand Oaks, California: Pine Forge Press (a Sage Publications company).

DeMaio, T.J. (1984). Social desirability and survey measurement: a review. In *Surveying Subjective Phenomena,* C.F. Turner and E. Martin, eds. New York: Russell Sage, pp. 257–282.

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

_____ (1983). Mail and other self-administered questionnaires. In *Handbook of Survey Research,* P. Rossi, R.A. Wright and B.A. Anderson, eds. New York: Academic Press, pp. 359–377.

_____ (1991). The design and administration of mail surveys. *Annual Review of Sociology*, vol. 17, pp. 225-249.

_____ (2000). Mail and Internet Surveys: *The Tailored Design Method*. New York: John Wiley and Sons.

Eisenhower, D., N.A. Mathiowetz and D. Morganstein (1991). Recall error: sources and bias reduction techniques. In *Measurement Errors in Surveys,* P. Biemer and others, eds. New York: John Wiley and Sons, pp.127–144.

Feindt, P., I. Schreiner and J. Bushery (1997). Reinterview: a tool for survey quality management. In *Proceedings of the Section on Survey Research Methods*. Alexandria, Virginia: American Statistical Association, pp. 105–110.

Forsman, G., and I. Schreiner (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*, P. Biemer and others, eds. New York: John Wiley and Sons, pp. 279–302.

Fowler, F.J. (1991). Reducing interviewer-related error through interviewer training, supervision and other means. In *Measurement Errors in Surveys*, P. Biemer and others, eds. New York: John Wiley and Sons, pp. 259–275.

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

_____ , and L.J. Magilavy (1986). Measuring and explaining interviewer effects. *Public Opinion Quarterly*, vol. 50, pp. 251–256.

Hastie, R, and D. Carlston (1980). Theoretical issues in person memory. In *Person Memory: The Cognitive Basis of Social Perception*, R. Hastie and others, eds. Hillsdale, New Jersey: Lawrence Erlbaum, pp. 1–53.

Hill, D.H. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics*, vol. 10, No. 4, pp. 359–380.

Huang, H. (1993). *Report on SIPP Recall Length Study*. Internal United States Bureau of the Census, Washington, D.C.

Institute for Resource Development/Macro Systems, Inc. (1990). *An Assessment of DHS-1 Data Quality*. Demographic and Health Surveys Methodological Reports, No. 1. Columbia, Maryland: Institute for Resource Development/Macro Systems, Inc.

Jenkins, C., and D. Dillman (1997). Towards a theory of self-administered questionnaire design. In *Survey Measurement and Process Quality*, L. Lyberg and others, eds. New York: John Wiley and Sons, pp. 165–196.

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

Kalton, G., D. Kasprzyk and D.B. McMillen (1989). Non-sampling errors in panel surveys. In *Panel Surveys*, D. Kasprzyk and others, eds. New York: John Wiley and Sons, pp. 249–270.

Kalton, G., D. McMillen and D. Kazprzyk (1986). Non-sampling error issues in SIPP. In *Proceedings of the Bureau of the Census Second Annual Research Conference*. Washington, D.C., pp.147-164.

Kantorowitz, M. (1992). *Methodological Issues in Family Expenditure Surveys*, Vitoria-Gasters, autonomous community of Euskadi: Euskal Estatistika-Erakundea, Instituto Vasco de Estadistica.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, vol. 57, pp. 92–115.

Lyberg, L., and D. Kasprzyk (1991). Data Collection Methods and Measurement Errors: An Overview. In *Measurement Errors in Surveys*, P. Biemer and others, eds. New York: John Wiley and Sons, pp.237–258.

_____ , P. Biemer, M. Collins, E.D. DeLeeuw, C. Dippo, N. Schwartz and D. Trewin (1997). In *Survey Measurement and Process Quality*. New York: John Wiley and Sons.

Marquis, K.H., and C.F. Cannell (1971). Effects of some experimental techniques on reporting in the health interview. In *Vital and Health Statistics*, Washington, D.C.: Public Health Service, Series 2 (Data Evaluation and Methods Research), No. 41.

_____ , and J.C. Moore (1990). Measurement errors in SIPP program reports. In *Proceedings of the Bureau of the Census 1990 Annual Research Conference*. Washington, D.C., pp. 721–745.

Mathiowetz, N. (2000). The effect of length of recall on the quality of survey data. In *Proceedings of the 4th International Conference on Methodological Issues in Official Statistics*. Stockholm: Statistics Sweden. Available from http://www.scb.se/Grupp/Omscb/_Dokument/Mathiowetz.pdf (Accessed 3 June 2004).

_____ , and K. McGonagle (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics*, vol. 16, pp. 401–418.

Neter, J. (1970). Measurement errors in reports of consumer expenditures. *Journal of Marketing Research*, vol. VII, pp. 11-25.

_____ , and J. Waksberg (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, vol. 59, pp. 8–55.

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

Oksenberg, L., C. Cannell and S. Blixt (1996). Analysis of interviewer and respondent behavior in the household survey. *National Medical Expenditure Survey Methods*, 7. Rockville, Maryland: Agency for Health Care and Policy Research, Public Health Service.

O'Muircheartaigh, C. (1982). *Methodology of the Response Errors Project*. WFS Scientific Reports, No. 28. Voorburg, Netherlands: International Statistical Institute.

_____ (1984a). *The Magnitude and Pattern of Response Variance in the Lesotho Fertility Survey*. WFS Scientific Reports, No. 70. Voorburg, Netherlands: International Statistical Institute.

_____ (1984b). *The Magnitude and Pattern of Response Variance in the Peru Fertility Survey*. WFS Scientific Reports, No. 45. Voorburg, Netherlands: International Statistical Institute.

Schreiner, I., K. Pennie and J. Newbrough (1988). Interviewer falsification in Census Bureau Surveys. In *Proceedings of the Section on Survey Research Methods*. Alexandria, Virginia: American Statistical Association, pp. 491–496.

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

Schwarz, N. (1997). Questionnaire design: the rocky road from concepts to answers. In *Survey Measurement and Process Quality*, L. Lyberg and others, eds. New York: John Wiley and Sons, pp. 29–46.

_____ , R.M. Groves and H. Schuman (1995). *Survey Methods*. Survey Methodology Program Working Paper Series. Ann Arbor, Michigan, Institute for Survey Research, University of Michigan.

_____, and H. Hippler (1991). Response alternatives: the impact of their choice and presentation order. In *Measurement Errors in Surveys*, P. Biemer and others, eds. New York: John Wiley and Sons, pp. 41–56.

_____, and S. Sudman (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco, California: Jossey-Bass.

Silberstein, A., and S. Scott (1991). Expenditure diary surveys and their associated errors. In *Measurement Errors in Surveys*, P. Biemer and others, eds. New York: John Wiley and Sons, pp. 303-326.

Singh, S. (1987). Evaluation of data quality. In *The World Fertility Survey: An Assessment*, J. Cleland and C. Scott, eds. New York: Oxford University Press, pp. 618-643.

Sirken, M. and others (1999). *Cognition and Survey Research*. New York: John Wiley and Sons.

Stanton, C., N. Abderrahim and K. Hill (1997). *DHS Maternal Mortality Indicators: An Assessment of Data Quality and Implications for Data Use*. Demographic and Health Surveys Analytical Report, No. 4. Calverton, Maryland: Macro International, Inc.

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

_____, and others (1977). Modest expectations: the effect of interviewers' prior expectations on response. *Sociological Methods and Research*, vol. 6, No. 2, pp. 171–182.

Tucker, C. (1997). Methodological issues surrounding the application of cognitive psychology in survey research. *Bulletin of Sociological Methodology*, vol. 55, pp.67–92.

United Nations (1982). *National Household Survey Capability Programme: Non-sampling Errors in Household Surveys: Sources, Assessment, and Control: Preliminary Version* DP/UN/INT-81-041/2. New York: United Nations Department of Technical Co-operation for Development and Statistical Office.

United States Bureau of the Census (1985). *Evaluating Censuses of Population and Housing*. Statistical Training Document. Washington, D.C. ISP-TR-5.

_____ (1998). *Survey of Income and Program Participation (SIPP) Quality Profile*, 3rd ed. Washington, D.C.: United States Department of Commerce.

United States Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Error in Surveys*, Statistical Policy working Paper, No. 31. Washington, D.C.: United States Office of Management and Budget. Available from http://www.fcsm.gov (accessed 14 May 2004).

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

Vaessen, M. and others (1987). Translation of questionnaires into local languages. In *The World Fertility Survey: An Assessment*, J. Cleland and C. Scott, eds. New York: Oxford University Press, pp.173-191.

Weiss, C. (1968), Validity of welfare mothers' interview response. *Public Opinion Quarterly,* vol. 32, pp. 622–633.

Westoff, C., N. Goldman and L. Moreno (1990). *Dominican Republic Experimental Study: An Evaluation of Fertility and Child Health Information*. Princeton, New Jersey: Office of Population Research; and Columbia, Maryland: Institute for Resource Development/Macro Systems, Inc.

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

_____ , P. Royston and D. Bercini (1991). The use of verbal report methods in the development and testing of survey questions. In *Applied Cognitive Psychology*, vol. 5, pp. 251-267.

Woltman, H.F., and J.B.Bushery (1977). *Update of the National Crime Survey Panel Bias Study*. Internal United States Bureau of the Census report, Washington, D.C.