6

Measurement Error in Surveys of the Low-Income Population

Nancy A. Mathiowetz, Charlie Brown, and John Bound

The measurement of the characteristics and behavioral experience among members of the low-income and welfare populations offers particular challenges with respect to reducing various sources of response error. For many of the substantive areas of interest, the behavioral experience of the welfare populations is complex, unstable, and highly variable over time. As the behavioral experience of respondents increases in complexity, so do the cognitive demands of a survey interview. Contrast the task of reporting employment and earnings for an individual continuously employed during the past calendar year with the response task of someone who has held three to four part-time jobs. Other questionnaire topics may request that the respondent report sensitive, threatening, socially undesirable, or perhaps illegal behavior. From both a cognitive and social psychological perspective, there is ample opportunity for the introduction of error in the reporting of the events and behaviors of primary interest in understanding the impacts of welfare reform.

This paper provides an introduction to these sources of measurement error and examines two theoretical frameworks for understanding the various sources of error. The empirical literature concerning the quality of responses for reports of earnings, transfer income, employment and unemployment, and sensitive behaviors is examined, to identify those items most likely to be subjected to response error among the welfare population. The paper concludes with suggestions for attempting to reduce the various sources of error through alternative questionnaire and survey design.

SOURCES OF ERROR IN THE SURVEY PROCESS

The various disciplines that embrace the survey method, including statistics, psychology, sociology, and economics, share a common concern with the weakness of the measurement process, the degree to which survey results deviate from "those that are the true reflections of the population" (Groves, 1989). The disciplines vary in the terminology used to describe error as well as their emphasis on understanding the impact of measurement error on analyses or the reduction of the various sources of error. The existence of these terminological differences and our desire to limit the focus of this research to *measurement error* suggests that a brief commentary on the various conceptual frameworks may aid in defining our interests unambiguously.

One common conceptual framework is that of mean squared error, the sum of the variance and the square of the bias. Variance is the measure of the variable error associated with a particular implementation of a survey; inherent in the notion of variable error is the fundamental requirement of replication, whether over units of observation (sample units), questions, or interviewers. Bias, as used here, is defined as the type of error that affects all implementations of a survey design, a constant error, within a defined set of essential survey conditions (Hansen et al., 1961). For example, the use of a single question to obtain total family income in the Current Population Survey (CPS) has been shown to underestimate annual income by approximately 20 percent (U.S. Bureau of the Census, 1979); this consistent underestimate would be considered the extent of the bias related to a particular question for a given survey design.

Another conceptual framework focuses on errors of observation as compared to errors of nonobservation (Kish, 1965). Errors of observation refer to the degree to which individual responses deviate from the true value for the measure of interest; as defined, they are the errors of interest for this research, to be referred to as measurement errors. Observational errors can arise from any of the elements directly engaged in the measurement process, including the questionnaire, the respondent, and the interviewer, as well as the characteristics that define the measurement process (e.g., the mode and method of data collection). Errors of nonobservation refer to errors related to the lack of measurement for some portion of the sample and can be classified as arising from three sources, coverage: nonresponse (both unit and item nonresponse), and sampling. Errors of nonobservation are the focus of other papers presented in this volume (see, for example, Groves and Couper, this volume).

Questionnaire as Source of Measurement Error

Ideally a question will convey to the respondent the meaning of interest to the researcher. However, several linguistic, structural, and environmental factors affect the interpretation of the question by the respondent. These factors include

the specific question wording, the structure of each question (open versus closed), and the order in which the questions are presented. Question wording is often seen as one of the major problems in survey research; although one can standardize the language read by the respondent or the interviewer, standardizing the language does not imply standardization of the meaning. In addition, a respondent's perception of the intent or meaning of a question can be shaped by the sponsorship of the survey, the overall topic of the questionnaire, or the environment more immediate to the question of interest, such as the context of the previous question or set of questions or the specific response options associated with the question.

Respondent as Source of Measurement Error

Once the respondent comprehends the question, he or she must retrieve the relevant information from memory, make a judgment as to whether the retrieved information matches the requested information, and communicate a response. The retrieval process is potentially fraught with error, including errors of omission and commission. As part of the communication of the response, the respondent must determine whether he or she wishes to reveal the information. Survey instruments often ask questions about socially and personally sensitive topics. It is widely believed, and well documented, that such questions elicit patterns of underreporting (for socially undesirable behaviors and attitudes) as well as overreporting (for socially desirable behaviors and attitudes).

Interviewers as Sources of Measurement Error

For interviewer-administered questionnaires, interviewers may affect the measurement processes in one of several ways, including:

- Failure to read the question as written;
- Variation in interviewers' ability to perform the other tasks associated with interviewing, for example, probing insufficient responses, selecting appropriate respondents, or recording information provided by the respondent; and
- Demographic and socioeconomic characteristics as well as voice characteristics that influence the behavior and responses provided by the respondent.

The first two factors contribute to measurement error from a cognitive or psycholinguistic perspective in that different respondents are exposed to different stimuli; thus variation in responses is, in part, a function of the variation in stimuli. All three factors suggest that interviewer effects contribute via an increase in variable error across interviewers. If all interviewers erred in the same direction (or their characteristics resulted in errors of the same direction and magnitude), interviewer bias would result. For the most part, the literature indicates that among

well-trained interviewing staff, interviewer error contributes to the overall variance of estimates as opposed to resulting in biased estimates (Lyberg and Kasprzyk, 1991).

Other Essential Survey Conditions as Sources of Measurement Error

Any data collection effort involves decisions concerning the features that define the overall design of the survey, here referred to as the essential survey conditions. In addition to the sample design and the wording of individual questions and response options, these decisions include:

- Whether to use interviewers or to collect information via some form of self-administered questionnaire;
- The means for selecting and training interviewers (if applicable);
- The mode of data collection for interviewer administration (telephone versus face to face);
- The choice of respondent rule, including the extent to which the design permits the reporting of information by proxy respondents;
- The method of data collection (paper and pencil, computer assisted);
- The extent to which respondents are encouraged to reference records to respond to factual questions;
- Whether to contact respondents for a single interview (cross-sectional design) or follow respondents over time (longitudinal or panel design);
- For longitudinal designs, the frequency and periodicity of measurement;
- The identification of the organization for whom the data are collected; and
- The identification of the data collection organization.

No one design or set of design features is clearly superior with respect to overall data quality. For example, as noted, interviewer variance is one source of variability that obviously can be eliminated through the use of a self-administered questionnaire. However, the use of an interviewer may aid in the measurement process by providing the respondent with clarifying information or by probing insufficient responses.

MEASUREMENT ERROR ASSOCIATED WITH AUTOBIOGRAPHICAL INFORMATION: THEORETICAL FRAMEWORK

Three distinct literatures provide the basis for the theoretical framework underlying investigations of measurement error in surveys. These theoretical foundations come from the fields of cognitive psychology, social psychology,

and to a lesser extent, social linguistics.[1] Although research concerning the existence, direction, magnitude as well as correlates of response error have provided insight into the factors associated with measurement error, there are few fundamental principles that inform either designers of data collection efforts or analysts of survey data as to the circumstances, either individual or design based, under which measurement error is most likely to be significant or not. Those tenets that appear to be robust across substantive areas are outlined in the following sections.

Cognitive Theory

Tourangeau (1984) as well as others (see Sudman et al., 1996, for a review) have categorized the survey question-and-answer process as a four-step process involving comprehension of the question, retrieval of information from memory, assessment of the correspondence between the retrieved information and the requested information, and communication. In addition, the encoding of information, a process outside the control of the survey interview, determines a priori whether the information of interest is available for the respondent to retrieve from long-term memory.

Comprehension of the interview question is the "point of entry" to the response process. Does the question convey the concept(s) of interest? Is there a shared meaning among the researcher, the interviewer, and the respondent with respect to each of the words as well as the question as a whole? The comprehension of the question involves not only knowledge of the particular words and phrases used in the questionnaire, but also the respondent's impression of the purpose of the interview, the context of the particular question, and the interviewer's behavior in the delivery of the question.

The use of simple, easily understood language is not sufficient for guaranteeing shared meaning among all respondents. Belson (1981) found that even simple terms were subject to misunderstanding. For example, Belson examined respondents' interpretation of the following question: "For how many hours do you usually watch television on a weekday? This includes evening viewing." He found that respondents varied in their interpretation of various terms such as "how many hours" (sometimes interpreted as requesting starting and stopping times of viewing), "you" (interpreted to include other family members), "usually," and "watch television" (interpreted to mean being in the room in which the television is on).

---

[1]Note that although statistical and economic theories provide the foundation for analysis of error-prone data, these disciplines provide little theoretical foundation for understanding the source of the measurement error nor the means for reducing measurement error. The discussion presented here will be limited to a review of cognitive and social psychological theories applicable to the measures of interest in understanding the welfare population.

Much of the measurement error literature has focused on the retrieval stage of the question-answering process, classifying the lack of reporting of an event as retrieval failure on the part of the respondent, comparing the characteristics of events that are reported to those that are not reported. One of the general tenets from this literature concerns the length of the recall period; the greater the length of the recall period, the greater the expected bias due to respondent retrieval and reporting error. This relationship has been supported by empirical data investigating the reporting of consumer expenditures and earnings (Neter and Waksberg, 1964); the reporting of hospitalizations, visits to physicians, and health conditions (e.g. Cannell et al., 1965); and reports of motor vehicle accidents (Cash and Moss, 1969), crime (Murphy and Cowan, 1976); and recreational activities (Gems et al., 1982). However, even within these studies, the findings with respect to the impact of the length of recall period on the quality of survey estimates are inconsistent. For example, Dodge (1970) found that length of recall was significant in the reporting of robberies but had no effect on the reporting of various other crimes, such as assaults, burglaries, and larcenies. Contrary to theoretically justified expectations, the literature also offers several examples in which the length of the recall period had no effect on the magnitude of response errors (see, for example, Mathiowetz and Duncan, 1988; Schaeffer, 1994). These more recent investigations point to the importance of the complexity of the behavioral experience over time, as opposed to simply the passage of time, as the factor most indicative of measurement error. This finding harkens back to theoretical discussions of the impact of interference on memory (Crowder, 1976).

Response errors associated with the length of the recall period typically are classified as either telescoping error, that is the tendency of the respondent to report events as occurring earlier (backward telescoping) or more recently (forward telescoping) than they actually occurred, or recall decay, the inability of the respondent to recall the relevant events occurring in the past (errors of omission). Forward telescoping is believed to dominate recall errors when the reference period for the questions is of short duration, while recall decay is more likely to have a major effect when the reference period is of long duration. In addition to the length of the recall period, the relative salience of the event affects the likelihood of either telescoping or memory decay. For example, events that are unique or that have a major impact on the respondent's life are less likely to be forgotten (error of omission) than less important events; however, the vividness of the event may lead respondents to recall the event as occurring more recently than is true (forward telescoping).

Another tenet rising from the collaborative efforts of cognitive psychologists and survey methodologists concerns the relationship between true behavioral experience and retrieval strategies undertaken by a respondent. Recent investigations suggest that the retrieval strategy undertaken by the respondent to provide a "count" of a behavior is a function of the true behavioral frequency. Research by Burton and Blair (1991) indicate that respondents choose to count events or items

(episodic enumeration) if the frequency of the event/item is low and they rely on estimation for more frequently occurring events. The point at which respondents switch from episodic counting to estimation varies by both the characteristics of the respondent and the characteristics of the event. As Sudman et al. (1996) note, "no studies have attempted to relate individual characteristics such as intelligence, education, or preference for cognitive complexity to the choice of counting or estimation, controlling for the number of events" (p. 201). Work by Menon (1993, 1994) suggests that it is not simply the true behavioral frequency that determines retrieval strategies, but also the degree of regularity and similarity among events. According to her hypotheses, those events that are both regular and similar (brushing teeth) require the least amount of cognitive effort to report, with respondents relying on retrieval of a rate to produce a response. Those events occurring irregularly require more cognitive effort on the part of the respondent.

The impact of different retrieval strategies with respect to the magnitude and direction of measurement error is not well understood; the limited evidence suggests that errors of estimation are often unbiased, although the variance about an estimate (e.g., mean value for the population) may be large. Episodic enumeration, however, appears to lead to biased estimates of the event or item of interest, with a tendency to be biased upward for short recall periods and downward for long recall periods.

A third tenet springing from this same literature concerns the salience or importance of the behavior to be retrieved. Sudman and Bradburn (1973) identify salient events as those that are unique or have continuing economic or social consequences for the respondent. Salience is hypothesized to affect the strength of the memory trace and subsequently, the effort involved in retrieving the information from long-term memory. The stronger the trace, the lower the effort needed to locate and retrieve the information. Cannell et al. (1965) report that those events judged to be important to the individual were reported more completely and accurately than other events. Mathiowetz (1986) found that short spells of unemployment were less likely to be reported than longer (i.e., more salient) spells.

The last maxim concerns the impact of interference related to the occurrence of similar events over the respondent's life or during the reference period of interest. Classical interference and information-processing theories suggest that as the number of similar or related events occurring to an individual increases, the probability of recalling any one of those events declines. An individual may lose the ability to distinguish between related events, resulting in an increase in the rate of errors or omission. Inaccuracy concerning the details of any one event also may increase as the respondent makes use of general knowledge or impressions concerning a class of events for reconstructing the specifics of a particular occurrence. Interference theory suggests that "forgetting" is a function of both the number and temporal pattern of related events in long-term memory. In addition,

we would speculate that interference also contributes to the misreporting of information, for example, the reporting of the receipt of Medicare benefits rather than Medicaid benefits.

Social Psychology: The Issue of Social Desirability

In addition to asking respondents to perform the difficult task of retrieving complex information from long-term memory, survey instruments often ask questions about socially and personally sensitive topics. Some topics are deemed, by social consensus, to be too sensitive to discuss in "polite" society. This was a much shorter list in the 1990s than in the 1950s, but most would agree that topics such as sexual practices, impotence, and bodily functions fall within this classification. Some (e.g., Tourangeau et al., 2000) hypothesize that questions concerning income also fall within this category. Other questions may concern topics that have strong positive or negative normative responses (e.g., voting, the use of pugnacious terms with respect to racial or ethnic groups) or for which there may be criminal retribution (e.g., use of illicit drugs, child abuse).

The sensitivity of the behavior or attitude of interest may affect both the encoding of the information as well as the retrieval and reporting of the material; little of the survey methodological research has addressed the point at which the distortion occurs with respect to the reporting of sensitive material. Even if the respondent is able to retrieve accurate information concerning the behavior of interest, he or she may choose to edit this information at the response formation stage as a means to reduce the costs, ranging from embarrassment to potential negative consequences beyond the interview situation, associated with revealing the information.

Applicability of Findings to the Measurement of Economic Phenomena

One of the problems in drawing inferences from other substantive fields to that of economic phenomena is the difference in the nature of the measures of interest. Much of the assessment of the quality of household-based survey reports concerns the reporting of discrete behaviors; many of the economic measures that are the subject of inquiry with respect to the measurement of the welfare population are not necessarily discrete behaviors or even phenomena that can be linked to a discrete memory. Some of the phenomena of interest could be considered trait phenomena. Let's consider the reporting of occupation. We speculate that the cognitive process by which one formulates a response to a query concerning current occupation is different from the process related to reporting the number of doctor visits during the past year.

For other economic phenomena, we speculate that individual differences in the approach to formulating a response impact the magnitude and direction of error associated with the measurement process. Consider the reporting of current

earnings related to employment. For some respondents, the request to report current earnings requires little cognitive effort—it may be almost an automatic response. For these individuals, wages may be considered a characteristic of their self-identity, a trait related to how they define themselves. For other individuals, the request for information concerning current wages may require the retrieval of information from a discrete episode (the last paycheck), the retrieval of a recent report of the information (the reporting of wages in an application for a credit card), or the construction of an estimate at the time of the query based on the retrieval of information relevant to the request.

Given both the theoretical and empirical research conducted within multiple branches of psychology and survey methodology, what would we anticipate are the patterns of measurement error for various economic measures? The response to that question is a function of how the respondent's task is formulated and the very nature of the phenomena of interest. For example, asking a respondent to provide an estimate of the number of weeks of unemployment during the past year is quite different from the task of asking the respondent to report the starting and stopping dates of each unemployment spell for the past year. For individuals in a steady state (constant employment or unemployment), neither task could be considered a difficult cognitive process. For these individuals, employment or unemployment is not a discrete event but rather may become encoded in memory as a trait that defines the respondent. However, for the individual with sporadic spells of unemployment throughout the year, the response formulation process most likely would differ for the two questions. Although the response formulation process for the former task permits an estimation strategy on the part of the respondent, the latter requires the retrieval of discrete periods of unemployment. For the reporting of these discrete events, we would hypothesize that patterns of response error evident in the reporting of events in other substantive fields would be observed. With respect to social desirability, we would anticipate patterns similar to those evident in other types of behaviors: overreporting of socially desirable behaviors and underreporting of socially undesirable behaviors.

Measurement Error in Household Reports of Income

As noted by Moore et al. (1999), the reporting of income by household respondents in many surveys can be characterized as a two-step process: the first involving the correct enumeration of sources of household income and the second, the accurate reporting of the amount of the income for the specific source. They find that response error in the reporting of various sources and amounts of income may be due to a large extent to cognitive factors, such as "definitional issues, recall and salience problems, confusion, and sensitivity" (p. 155). We return to these cognitive factors when considering alternative means for reducing measurement error in surveys of the low-income population.

Earnings

Empirical evaluations of household-reported earnings information include the assessment of annual earnings, usual earnings (with respect to a specific pay period), most recent earnings, and hourly wage rates. These studies rely on various sources of validation data, including the use of employers' records, administrative records, and respondents' reports for the same reference period reported at two different times.

With respect to reports of annual earnings, mean estimates appear to be subject to relatively small levels of response error, although absolute differences indicate significant overreporting and underreporting at the individual level. For example, Borus (1970) focused on survey responses of residents in low-income census tracts in Fort Wayne, Indiana. The study examined two alternative approaches to questions concerning annual earnings: (1) the use of two relatively broad questions concerning earnings, and (2) a detailed set of questions concerning work histories. Responses to survey questions were compared to data obtained from the Indiana Employment Security Division for employment earnings covered by the Indiana Unemployment Insurance Act. Borus found that the mean error in reports of annual earnings was small and insignificant for both sets of questions; however, more than 10 percent of the respondents misreported annual earnings by $1,000 (based on a mean of $2,500). Among poor persons with no college education, Borus found that the broad questions resulted in more accurate data than the work history questions.

Smith (1997) examined the reports of earnings data among individuals eligible to participate in federal training programs. Similar to the work by Borus (1970), Smith compared the reports based on direct questions concerning annual earnings to those responses based on summing the report of earnings for individual jobs. The decomposition approach, that is, the reporting of earnings associated with individual jobs, led to higher reports of annual earnings, attributed to both an increase in the reporting of number of hours worked as well as an increase in the reporting of irregular earnings (overtime, tips, and commissions). Comparisons with administrative data for these individuals led Smith to conclude that the estimates based on adding up earnings across jobs led to overreporting, rather than more complete reporting.[2]

Duncan and Hill (1985) sampled employees from a single establishment and compared reports of annual earnings with information obtained from the employer's records. The nature of the sample, employed persons, limits our ability

[2]An alternative interpretation of the findings might suggest that the decomposition approach was more accurate and that the apparent overestimation, when compared to administrative records, is because of underreporting of income in the administrative records rather than overreporting of earnings using the decomposition method.

to draw inferences from their work to the low-income population. Respondents were interviewed in 1983 and requested to report earnings and employment-related measures for calendar years 1981 and 1982. For neither year was the mean of the sample difference between household-based reports and company records statistically significant (8.5 percent and 7 percent of the mean, respectively), although the absolute differences for each year indicate significant underreporting and overreporting. Comparison of measures of change in annual earnings based on the household report and the employer records indicate no difference; interview reports of absolute change averaged $2,992 (or 13 percent) compared to the employer-based estimate of $3,399 (or 17 percent).

Although the findings noted are based on small samples drawn from either a single geographic area (Borus) or a single firm (Duncan and Hill), the results parallel the findings from empirical research comprised of nationally representative samples. Bound and Krueger (1991) examined error in annual earnings as reported in the March, 1978 CPS. Although the error was distributed around approximately a zero mean for both men and women, the magnitude of the error was substantial.

In addition to examining bias in mean estimates, the studies by Duncan and Hill and Bound and Krueger examined the relationship between measurement error and true earnings. Both studies indicate a significant negative relationship between error in reports of annual earnings and the true value of annual earnings. Similar to Duncan and Hill (1985), Bound and Krueger (1991) report positive autocorrelation (.4 for men and .1 for women) between errors in CPS-reported earnings for the 2 years of interest, 1976 and 1977.

Both Duncan and Hill (1985) and Bound and Krueger (1991) explore the implications of measurement error for earnings models. Duncan and Hill's model relates the natural logarithm of annual earnings to three measures of human capital investment: education, work experience prior to current employer, and tenure with current employer, using both the error-ridden self-reported measure of annual earnings and the record-based measure as the left-hand-side variable. A comparison of the ordinary least squares parameter estimates based on the two dependent variables suggests that measurement error in the dependent variable has a sizable impact on the parameter estimates. For example, estimates of the effects of tenure on earnings based on interview data were 25 percent lower than the effects based on record earnings data. Although the correlation between error in reports of earnings and error in reports of tenure was small (.05) and insignificant, the correlation between error in reports of earnings and *actual* tenure was quite strong (–.23) and highly significant, leading to attenuation in the estimated effects of tenure on earnings based on interview information.

Bound and Krueger (1991) also explore the ramifications of an error-ridden left-hand-side variable by regressing error in reports of earnings with a number of human capital and demographic factors, including education, age, race, marital status, region, and standard metropolitan statistical area (SMSA). Similar to

Duncan and Hill, the model attempts to quantify the extent to which the correlation between measurement error in the dependent variable and right-hand-side variables biases the estimates of the parameters. However, in contrast to Duncan and Hill, Bound and Krueger conclude that mismeasurement of earnings leads to little bias when CPS-reported earnings are on the left-hand side of the equation.

The reporting of annual earnings within the context of a survey is most likely aided by the number of times the respondent has retrieved and reported the information. For some members of the population, we contend that the memory for one's annual earnings is reinforced throughout the calendar year, for example, in the preparation of federal and state taxes or the completion of applications for credit cards and loans. To the extent that these requests have motivated the respondent to determine and report an accurate figure, such information should be encoded in the respondent's memory. Subsequent survey requests therefore should be "routine" in contrast to many of the types of questions posed to a survey respondent. Hence we would hypothesize that response error in such situations would result from retrieval of the wrong information (e.g., annual earnings for calendar year 1996 rather than 1997; net rather than gross earnings), social desirability issues (e.g., overreporting among persons with low earnings related to presentation of self to the interviewer), or privacy concerns, which may lead to either misreporting or item nonresponse.

Although the limited literature on the reporting of earnings among the low-income population indicates a high correlation between record and reported earnings (Halsey, 1978), we hypothesize that for some members of the population—such as low-income individuals for whom there are fewer opportunities to retrieve and report annual earnings information—a survey request would not be routine and may require very different response strategies than for respondents who have regular opportunities to report their annual earnings. Only two studies cited here, Borus (1970) and Smith (1997), compared alternative approaches to the request for earnings information among the low-income population. Borus found that the broad-based question approach led to lower levels of response error than a work history approach and Smith concluded that a decomposition approach led to an overestimation of annual earnings. The empirical results of Borus and Smith suggest, in contrast to theoretical expectations, that among the lower income populations, the use of broad questions may result in more accurate reports of income than detailed questions related to each job. Despite these findings, we speculate that for the low income population, those with loose ties to the labor force, or those for whom the retrieval of earnings information requires separate estimates for multiple jobs, the use of a decomposition approach or some type of estimation approach may be beneficial and warrants additional research.

In contrast to the task of reporting annual earnings, the survey request to report weekly earnings, most recent earnings, or usual earnings is most likely a relatively unique request and one that may involve the attempted retrieval of information that may not have been encoded by the respondent, the retrieval of

information that has not been accessed by the respondent before, or the calculation of an estimate "on the spot." To the extent that the survey request matches the usual reference period for earnings (e.g., weekly pay), we would anticipate that requests for the most recent period may be well reported. In contrast, we would anticipate that requests for earnings in any metric apart from a well-rehearsed metric would lead to significant differences between household reports and validation data.

A small set of studies examined the correlation between weekly or monthly earnings as reported by workers and their employer's reports (Keating et al., 1950; Hardin and Hershey, 1960; Borus, 1966; Dreher, 1977). Two of these studies focus on the population of particular interest, unemployed workers (Keating et al., 1950) and training program participants (Borus, 1966). All four studies report correlations between the employee's report and the employer's records of .90 or higher. Mean reports by workers are close to record values, with modest overreporting in some studies and underreporting in others. For example, Borus (1966) reports a high correlation (.95) between household and employer's records of weekly earnings, small mean absolute deviations between the two sources, and equal amounts of overreporting and underreporting.

Carstensen and Woltman (1979), in a study among the general population, compared worker and employer reports, based on a supplement to the January, 1977 CPS. Their survey instruments allowed both workers and employers to report earnings in whatever time unit they preferred (e.g., annually, monthly, weekly, hourly). Comparisons were limited to those reports for which the respondent and the employer reported earnings using the same metric. When earnings were reported by both worker and employer on a weekly basis, workers underreported their earnings by 6 percent; but when both reported on a monthly basis, workers overreported by 10 percent.

Rodgers et al. (1993)[3] report correlations of .60 and .46 between household reports and company records for the most recent and usual pay, respectively, in contrast to a correlation of .79 for reports of annual earnings. In addition, they calculated an hourly wage rate from the respondents' reports of annual, most recent, and usual earnings and hours and compared that hourly rate to the rate as reported by the employer; error in the reported hours for each respective time period therefore contributes to noise in the hourly wage rate. Similar to the findings for earnings, correlation between the employer's records and self-reports were highest when based on annual earnings and hours (.61) and significantly lower when based on most recent earnings and hours and usual earnings and hours (.38 and .24, respectively).

[3]Based on the Panel Study of Income Dynamics validation study, a survey conducted among a sample of employees at a single establishment and comparing their responses to those obtained from company records. The data from the first wave of this two-wave study were the basis for the study reported by Duncan and Hill (1985).

Hourly wages calculated from the CPS-reported earnings and hours compared to employers' records indicate a small but significant rate of underreporting, which may be due to an overreporting of hours worked, an underreporting of annual earnings, or a combination of the two (Mellow and Sider, 1983). Similar to Duncan and Hill (1985), Mellow and Sider examined the impact of measurement error in wage equations; they concluded that the structure of the wage determination process model was unaffected by the use of respondent- or employer-based information, although the overall fit of the model was somewhat higher with employer-reported wage information.

As noted earlier, one of the shortfalls with the empirical investigations concerning the reporting of earnings is the lack of studies targeted at those for whom the reporting task is most difficult—those with multiple jobs or sporadic employment. Although the empirical findings suggest that annual earnings are reported more accurately than earnings for other periods of time, the opposite may be true among those for whom annual earnings are highly variable and the result of complex employment patterns.

One of the major concerns with respect to earnings questions in surveys of Temporary Assistance for Needy Families (TANF) leavers is the reference period of interest. Many of the surveys request that respondents report earnings for reference periods that may be of little salience to the respondent or for which the determination of the earnings is quite complex. For example, questions often focus on the month in which the respondent left welfare (which may have been several months prior to the interview) or the 6 month period prior to exiting welfare. The movement off welfare support would probably be regarded as a significant and salient event and therefore be well reported. However, asking the respondent to reconstruct a reference period prior to the month of exiting welfare is most likely a cognitively difficult task. For example, consider the following question:

- During the six months you were on welfare before you got off in MONTH, did you ever have a job which paid you money?

For this question, the reference period of interest is ambiguous. For example, if the respondent exited welfare support in November 1999, is the 6-month period of interest defined as May 1, 1999, through October 31, 1999, or is the respondent to include the month in which he or she exited welfare as part of the reference period, in this case, June 1999-November 1999? If analytic interest lies in understanding a definitive period prior to exiting welfare, then the questionnaire should explicitly state this period to the respondent (e.g., "In the 6 months prior to going off welfare, that is, between May 1 and October 31, 1999") as well as encourage the respondent to use a calendar or other records to aid recall. The use of a calendar may be of particular importance when the reference period spans 2 calendar years. If the analytic interest lies in a more diffuse measure of employ-

ment in some period prior to exiting welfare, a rewording of the question so as to not imply precision about a particular 6 months may be more appropriate.

TRANSFER PROGRAM INCOME AND CHILD SUPPORT

For most surveys, the reporting of transfer program income is a two-stage process in which respondents first report recipiency (or not) of a particular form of income and then, among those who report recipiency, the amount of the income. One shortcoming of many studies that assess response error associated with transfer program income is the design of the study, in which the sample for the study is drawn from those known to be participants in the program. Responses elicited from respondents then are verified with administrative data. Retrospective or reverse record check studies limit the assessment of response error, with respect to recipiency, to determining the rate of underreporting; prospective or forward record check studies that only verify positive recipiency responses are similarly flawed because by design they limit the assessment of response error only to overreports. In contrast, a "full" design permits the verification of both positive and negative recipiency responses and includes in the sample a full array of respondents. Validation studies that sample from the general population and link all respondents, regardless of response, to the administrative record of interest represent full study designs.

We focus our attention first on reporting of receipt of a particular transfer program. Among full design studies, there does appear to be a tendency for respondents to underreport receipt, although there are also examples of overreporting recipiency status. For example, Oberheu and Ono (1975) report a low correspondence between administrative records and household report for receipt of Aid to Families with Dependent Children (AFDC)—monthly and annual—and food stamps (disagreement rates exceeding 20 percent), but relatively low net rates of underreporting and overreporting. Underreporting of the receipt of general assistance as reported in two studies is less than 10 percent (e.g., David, 1962). In a study reported by Marquis and Moore (1990), respondents were asked to report recipiency status for 8 months (in two successive waves of Survey of Income and Program Participation [SIPP] interviews). Although Marquis and Moore report a low error rate of approximately 1 percent to 2 percent, the error rate among true recipients is significant, in the direction of underreporting. For example, among those receiving AFDC, respondents failed to report receipt in 49 percent of the person-months. Underreporting rates were lowest among Old-Age and Survivors Insurance and Disability Insurance (OASDI) beneficiaries, for which approximately 5 percent of the person-months of recipiency were not reported by the household respondents. The mean rates of participation based on the two sources differed by less than 1 percentage point for all income types. However, because some of these programs are so rare, small absolute biases mask high rates of relative underreporting among true participants, ranging from

+1 percent for OASDI recipiency to nearly 40 percent for AFDC recipiency. In a followup study, Moore et al. (1996) compared underreporting rates of known recipients to overreporting rates for known nonrecipients and found underreporting rates to be much higher than the rate of false positives by nonrecipients. They also note that underreporting on the part of known recipients tends to be due to failure to *ever* report receipt of a particular type of income rather than failure to report specific months of receipt.

In contrast, Yen and Nelson (1996) found a slight tendency among AFDC recipients to overreport receipt in any given month, such that estimates based on survey reports exceeded estimates based on records by approximately 1 percentage point. Oberheu and Ono (1975) also note a net overreporting for AFDC (annual) and food stamp recipiency (annual), of 8 percent and 6 percent, respectively. Although not investigated by these researchers, one possible explanation for apparent overreporting on the part of the respondent is confusion concerning the source of recipiency, resulting in an apparent overreporting of one program coupled with an underreporting of another program. Because many of the validity studies that use administrative records to confirm survey reports are limited to verification of one or two particular programs, most response error investigations have not addressed this problem.

Errors in the reporting of recipiency for any given month may be attributable to misdating the beginning and end points of a spell, as opposed to an error of omission or confusion concerning the source of support. The "seam effect" refers to a particular type of response error resulting from the misdating of episodic information in panel data collection efforts (Hill, 1987). A seam effect is evident when a change in status (e.g., from receipt of AFDC to nonreceipt of AFDC) corresponds to the end of a reference period for Wave x and the beginning of a reference period for Wave x+1. For example, a respondent may report receipt of AFDC at the end of the first wave of interviewing; at the time of the second wave of interviewing, he or she reports that no one in the family has received such benefits for the entire reference period. Hence it appears (in the data) as if the change in status occurred on the day of the interview.

With respect to the direction and magnitude of estimates concerning the amount of the transfer, empirical investigations vary in their conclusions. Several studies report a significant underreporting of assistance amount (e.g., David, 1962; Livingston, 1969; Oberheu and Ono, 1975; Halsey, 1978) or significant differences between the survey and record reports (Grondin and Michaud, 1994). Other studies report little to no difference in the amount based on the survey and record reports. Hoaglin (1978) found no difference in median response error for welfare amounts and only small negative differences in the median estimates for monthly Social Security income. Goodreau et al. (1984) found that 65 percent of the respondents accurately report the amount of AFDC support; the survey report accounted for 96 percent of the actual amount of support. Although Halsey (1978) reported a net bias in the reporting of unemployment insurance amount of –50

percent, Dibbs et al. (1995) conclude that the average household report of unemployment benefits differed from the average true value by approximately 5 percent ($300 on a base of $5,600).

Schaeffer (1994) compared custodial parents' reports of support owed and support paid to court records among a sample of residents in the state of Wisconsin. The distribution of response errors indicated significant underreporting and overreporting of both the amount owed and the amount paid. The study also examined the factors contributing to the absolute level of errors in the reports of amounts owed and paid; the findings indicate that the complexity of the respondent's support experience had a substantial impact on the accuracy of the reports. Characteristics of the events (payments) were more important in predicting response error than characteristics of the respondent or factors related to memory decay. The analysis suggests two areas of research directed toward improving the reporting of child support payments: research related to improving the comprehension of the question (specifically clarifying and distinguishing child support from other transfer payments) and identifying respondents for whom the reporting process is difficult (e.g., use of a filter question) with follow-up questions specific to the behavioral experience.

Hours Worked

The number of empirical investigations concerning the quality of household reports of hours worked are few in number but consistent with respect to the findings. Regardless of whether the measure of interest is hours worked last week, annual work hours, usual hours worked, or hours associated with the previous or usual pay period, comparisons between company records and respondents' reports indicate an overestimate of the number of hours worked. We note that none of the empirical studies examined in the following text focuses specifically on the low-income or welfare populations.

Carstensen and Woltman (1979) assessed reports of "usual" hours worked per week. They found that compared to company reports, estimates of the mean usual hours worked were significantly overreported by household respondents: 37.1 hours versus 38.4 hours, respectively, a difference on average of 1.33 hours, or 3.6 percent of the usual hours worked. Similarly, Mellow and Sider (1983) report that the mean difference between the natural log of worker-reported hours and the natural log of employer-reported hours is positive (.039). Self-reports exceeded employer records by nearly 4 percent on average; however, for approximately 15 percent of the sample, the employer records exceeded the estimate provided by the respondent. A regression explaining the difference between the two sources indicates that professional and managerial workers were more likely to overestimate their hours, as were respondents with higher levels of education and nonwhite respondents. In contrast, female respondents tended to underreport usual hours worked.

Similar to their findings concerning the reporting of earnings, Rodgers et al. (1993) report that the correlation between self-reports and company records is higher for annual number of hours worked (.72) than for either reports of hours associated with the previous pay period (.61) or usual pay period (.61). Barron et al. (1997) report a high correlation between employers' records and respondents' reports of hours last week, .769. Measurement error in hours worked is not independent of the true value; as reported by Rodgers et al. (1993), the correlation between error in reports of hours worked and true values (company records) ranged from –.307 for annual hours worked in the calendar year immediately prior to the date of the interview to –.357 for hours associated with the previous pay period and –.368 for hours associated with usual pay period.

Examination of a standard econometric model with earnings as the left-hand-side variable and hours worked as one of the predictor variables indicates that the high correlation between the errors in reports of earnings and hours (ranging from .36 for annual measures to .54 for last pay period) seriously biases parameter estimates. For example, regressions of reported and company record annual earnings (log) on record or reported hours, age, education, and tenure with the company provide a useful illustration of the consequences of measurement error. Based on respondent reports of earnings and hours, the coefficient for hours (log hours) is less than 60 percent of the coefficient based on company records (.41 versus 1.016) while the coefficient for age is 50 percent larger in the model based on respondent reports. In addition, the fit of the model based on respondent reports is less than half that of the fit based on company records (R^2 of .352 versus .780).

Duncan and Hill (1985) compare the quality of reports of annual hours worked for two different reference periods, the prior calendar year and the calendar year ending 18 months prior to the interview. The quality of the household reports declines as a function of the length of the recall period, although the authors report significant overreporting for each of the two calendar years of interest. The average absolute error in reports of hours worked (157 hours) was nearly 10 percent of the mean annual hours worked for 1982 (μ=1,603) and nearly 12 percent (211 hours) of the mean for 1981 (μ=1,771). Comparisons of changes in hours worked reveal that although the simple differences calculated from two sources have similar averages, the absolute amount of change reported in the interview significantly exceeds that based on the record report.

In contrast to the findings with respect to annual earnings, we see both a bias in the population estimates as well as a bias in the individual reports of hours worked in the direction of overreporting. This finding persists across different approaches to measuring hours worked, regardless of whether the respondent is asked to report on hours worked last week (CPS) or account for the weeks worked last year, which then are converted to total hours worked during the year (Panel Study of Income Dynamics [PSID]). Whether this is a function of social desirability or whether it is related to the cognitive processes associated with

formulating a response to the questions measuring hours worked is something that can only be speculated on at this point. One means by which to attempt to repair the overreporting of hours worked is through the use of time-use diaries, where respondents are asked to account for the previous 24-hour period. Employing time-use diaries has been found to be an effective means for reducing response error associated with retrospective recall bias as well as bias associated with the overreporting of socially desirable behavior (Presser and Stinson, 1998).

Unemployment

In contrast to the small number of studies that assess the quality of household reports of hours worked, there are a number of studies that have examined the quality of unemployment reports. These studies encompass a variety of unemployment measures, including annual number of person-years of unemployment, weekly unemployment rate, occurrence and duration of specific unemployment spells, and total annual unemployment hours. Only one study reported in the literature, the PSID validation study (Duncan and Hill, 1985; Mathiowetz, 1986; Mathiowetz and Duncan, 1988), compares respondents' reports with validation data; the majority of the studies rely on comparisons of estimates based on alternative study designs or examine the consistency in reports of unemployment duration across rounds of data collection. In general, the findings suggest that retrospective reports of unemployment by household respondents underestimate unemployment, regardless of the unemployment measure of interest. Once again, however, these studies focus on the general population; hence our ability to draw inferences to the low income or welfare populations is limited.

The studies by Morganstern and Bartlett (1974), Horvath (1982), and Levine (1993) compare the contemporaneous rate of unemployment as produced by the monthly CPS to the rate resulting from retrospective reporting of unemployment during the previous calendar year.[4] The measures of interest vary from study to study; Morganstern and Bartlett focus on annual number of person-years of unemployment as compared to average estimates of weekly unemployment (Horvath) or an unemployment rate, as discussed by Levine. Regardless of the

[4]The CPS is collected each month from a probability sample of approximately 50,000 households; interviews are conducted during the week containing the 19th day of the month: respondents are questioned about labor force status for the previous week, Sunday through Saturday, which includes the 12th of the month. In this way, the data are considered the respondent's current employment status, with a fixed reference period for all respondents, regardless of which day of the week they are interviewed. In addition to the core set of questions concerning labor force participation and demographic characteristics, respondents interviewed in March of any year are asked a supplemental set of questions (hence the name March supplement) concerning income recipiency and amounts, weeks employed, unemployed and not in the labor force, and health insurance coverage for the previous calendar year.

measure of interest, the empirical findings from the three studies indicate that when compared to the contemporaneous measure, retrospective reports of labor force status result in an underestimate of the unemployment rate.

Across the three studies, the underreporting rate is significant and appears to be related to demographic characteristics of the individual. For example, Morganstern and Bartlett (1974) report discrepancy rates in the range of around 3 percent to 24 percent with the highest discrepancy rates among women (22 percent for black women; 24 percent for white women). Levine compared the contemporaneous and retrospective reports by age, race, and gender. He found the contemporaneous rates to be substantially higher relative to the retrospective reports for teenagers, regardless of race or sex, and for women. Across all of the years of the study, 1970-1988, the retrospective reports for white males, ages 20 to 59, were nearly identical to the contemporaneous reports.

Duncan and Hill (1985) found that the overall estimate of mean number of hours unemployed in years *t* and *t-1* based on employer reports and company records did not differ significantly. However, microlevel comparisons, reported as the average absolute difference between the two sources, were large relative to the average amount of unemployment in each year, but significant only for reports of unemployment occurring in 1982.

In addition to studies examining rates of unemployment, person-years of unemployment, or annual hours of unemployment, several empirical investigations have focused on spell-level information, examining reports of the specific spell and duration of the spell. Using the same data as presented in Duncan and Hill (1985), Mathiowetz and Duncan (1988) found that at the spell level, respondents failed to report more than 60 percent of the individual spells. Levine (1993) found that 35 percent to 60 percent of persons failed to report an unemployment spell one year after the event. In both studies, failure to report a spell of unemployment was related, in part, to the length of the unemployment spell; short spells of unemployment were subject to higher rates of underreporting.

The findings suggest (Poterba and Summers, 1984) that, similar to other types of discrete behaviors and events, the reporting of unemployment is subject to deterioration over time. However, the passage of time may not be the fundamental factor affecting the quality of the reports; rather the complexity of the behavioral experience over longer recall periods appears to be the source of increased response error. Both the microlevel comparisons as well as the comparisons of population estimates suggest that behavioral complexity interferes with the respondent's ability to accurately report unemployment for distant recall periods. Hence we see greater underreporting among population subgroups who traditionally have looser ties to the labor force (teenagers, women). Although longer spells of unemployment appear to be subject to lower levels of errors of omission, a finding that supports other empirical research with respect to the effects of salience, at least one study found that errors in reports of duration were associated negatively with the length of the spell. Whether this is indicative of an

error in cognition or an indication of reluctance to report extremely long spells of unemployment (social desirability) is unresolved.

Sensitive Questions: Drug Use, Abortions

A large body of methodological evidence indicates that embarrassing or socially undesirable behaviors are misreported in surveys (e.g., Bradburn, 1983). For example, comparisons between estimates of the number of abortions based on survey data from the National Survey of Family Growth (NSFG) and estimates based on data collected from abortion clinics suggest that fewer than half of all abortions are reported in the NSFG (Jones and Forrest, 1992). Similarly, comparisons of survey reports of cigarette smoking with sales figures indicates significant underreporting on the part of household respondents, with the rate of underreporting increasing over time, a finding attributed by the authors as a function of increasing social undesirability (Warner, 1978).

Although validation studies of reports of sensitive behaviors are rare, there is a growing body of empirical literature that examines reports of sensitive behaviors as a function of mode of data collection, method of data collection, question wording, and context (e.g., Tourangeau and Smith, 1996). These studies have examined the reporting of abortions, AIDS risk behaviors, use of illegal drugs, and alcohol consumption. The hypothesis for these studies is that, given the tendency to underreport sensitive or undesirable behavior, the method or combination of essential survey design features that yields the highest estimate is the "better" measurement approach.

Studies comparing self-administration to interviewer-administered questions (either face to face or telephone) indicate that self-administration of sensitive questions increases levels of reporting relative to administration of the same question by an interviewer. Increases in the level of behavior have been reported in self-administered surveys (using paper and pencil questionnaires) concerning abortions (London and Williams, 1990), alcohol consumption (Aquilino and LoSciuto, 1990), and drug use (Aquilino, 1994). Similar increases in the level of reporting sensitive behaviors have been reported when the comparisons focus on the difference between interviewer-administered questionnaires and computer-assisted self administration (CASI) questionnaires.

One of the major concerns with moving from an interviewer-administered questionnaire to self-administration is the problem of limiting participation to the literate population. Even among the literate population, the use of self-administered questionnaires presents problems with respect to following directions (e.g., skip patterns). The use of audio computer-assisted self-interviewing (ACASI) techniques circumvents both problems. The presentation of the questions in both written and auditory form (through headphones) preserves the privacy of a self-administered questionnaire without the restriction imposed by respondent lit-

eracy. The use of computers for the administration of the questionnaire eliminates two problems often seen in self-administered paper and pencil questionnaires—missing data and incorrectly followed skip patterns. A small but growing body of literature (e.g., O'Reilly et al., 1994; Tourangeau and Smith, 1996) finds that ACASI methods are acceptable to respondents and appear to improve the reporting of sensitive behaviors. Cynamon and Camburn (1992) found that using portable cassette players to administer questions (with the respondent recording answers on a paper form) also was effective in increasing reports of sensitive behaviors.

Methods for Reducing Measurement Error

As we consider means for reducing measurement error in surveys of the low-income population, we return to the theoretical frameworks that address the potential sources of error: those errors associated with problems of cognition and those resulting from issues associated with social desirability.

REPAIRS FOCUSING ON PROBLEM OF COGNITION

Comprehension

Of primary importance in constructing question items is to assure comprehension on the part of the respondent. Although the use of clear and easily understood language is a necessary step toward achieving that goal, simple language alone does not guarantee that the question is understood in the same manner by all respondents.

The literature examining comprehension problems in the design of income questions indicates that defining income constructs in a language easily understood by survey respondents is not easy (Moore et al., 1999). Terms that most researchers would consider to be well understood by respondents may suffer from differential comprehension. For example, Stinson (1997) found significant diversity with respect to respondents' interpretations of the term "total family income." Similarly, Bogen (1995) reported that respondents tend to omit sporadic self-employment and earnings from odd jobs or third or fourth jobs in their reports of income due to the respondents' interpretations of the term "income." These findings suggest the need for thorough testing of items among the population of interest to assess comprehension.

Comprehension of survey questions is affected by several factors, including the length of the question, the syntactical complexity, the degree to which the question includes instructions such as inclusion and exclusion clauses, and as the use of ambiguous terms. Consider, for example, the complexity of the following questions:

Example 1: Since your welfare benefits ended in (FINAL BENEFIT MONTH), did you take part for at least one month in any Adult Basic Education (ABE) classes for improving your basic reading and math skills, or General Education Development (GED) classes to help you prepare for the GED test, or classes to prepare for a regular high school diploma?

Example 2: In (PRIOR MONTH), did you have any children of your own living in the household? Please include any foster or adopted children. Also include any grandchildren living with you.

Example 3: Since (FINAL BENEFIT MONTH), have you worked for pay at a regular job at all? Please don't count unpaid work experience, but do include any paid jobs, including paid community service jobs or paid on-the-job training.

Each of these items is cognitively complex. The first question requires the respondent to process three separate categories of education, determine whether the conditional phrase "at least one month" applies only to the adult basic education classes or also to the GED and regular high school classes, and also attribute a reason for attending ABE ("improving reading and math skills") or GED classes. Separating example 1 into three simple items, prefaced by an introductory statement concerning types of education, would make the task more manageable for the respondent. Examples 2 and 3 suffer from the problem of providing an exclusion or inclusion (or in the case of example 3, both) clause after the question. Both would be improved by defining for the respondent what the question concerns and then asking the question, so that the last thing the respondent hears is the question. Example 2 may be improved by simply asking separate questions concerning own children, foster children, and grandchildren. Although questionnaire designers may be reluctant to add questions to an instrument for fear of longer administration times, we speculate that the administration of several well-designed short questions actually may be shorter than confusing compound questions that may require repeating or clarification.

With respect to question length, short questions are not always better. Cannell and colleagues (Cannell et al., 1977; Cannell et al., 1981) demonstrated that longer questions providing *redundant* information can lead to increased comprehension, in part because the longer question provides additional context for responding as well as longer time for the respondent to think about the question and formulate a response. On the other hand, longer questions that introduce new terms or become syntactically complex will result in lower levels of comprehension.

Comprehension can suffer from both lexical and structural ambiguities. For example, the sentence "John went to the bank" could be interpreted as John going to a financial institution or the side of a river. Lexical problems are inherent in a language in which words can have different interpretations. Although difficult to

fix, interpretation can be aided through context and the respondent's usual use of the word (in this case, most likely the financial institution interpretation). Note that when constructing a question, one must consider regional and cultural differences in language and avoid terms that lack a clearly defined lexical meaning (e.g., "welfare reform"). Structural ambiguities arise when the same word can be used as different parts of speech—for example, as both a verb or an adjective in the sentence "Flying planes can be dangerous." Structural ambiguities most often can be repaired through careful wording of the question.

Questionnaire designers often attempt to improve comprehension by grouping questions so as to provide a context for a set of items, writing explicit questions, and, if possible, writing closed-ended items in which the response categories may aid in the interpretation of the question by the respondent. In addition, tailoring questions to accommodate the language of specific population subgroups is feasible with computer-assisted interviewing systems.

Comprehension difficulties are best identified and repaired through the use of selected pretesting techniques such as cognitive interviewing or expert panel review (e.g., Presser and Blair, 1994; Forsyth and Lessler, 1991). Requesting respondents to paraphrase the question in their own words often provides insight into different interpretations of a question; similarly, the use of other cognitive interviewing techniques such as think-aloud interviews or the use of vignettes can be useful in identifying comprehension problems as well as offer possible alternative wording options for the questionnaire designer.

Retrieval

Many of the questions of interest in surveying the welfare population request that the respondent report on retrospective behavior, often for periods covering several years or more (e.g., year of first receipt of AFDC benefits). Some of these questions require that the respondent date events of interest, thus requiring episodic retrieval of a specific event. Other questions request that respondents provide a numeric estimate (e.g., earnings from work last month); in these cases the respondent may rely on episodic retrieval (e.g., the more recent pay-check), reconstruction, an estimation strategy, or a combination of retrieval strategies to provide a response. As noted earlier, response strategies are often a function of the behavioral complexity experienced by the respondent; however, the strategy used by the respondent can be affected by the wording of the question.

Although both responses based on episodic enumeration and estimation are subject to measurement error, the literature suggests that questions which direct the respondent toward episodic enumeration tend to suffer from errors of omissions (underreports) due to incomplete memory searches on the part of the respondent, whereas responses based on estimation strategies result in both inclusion and exclusion errors, resulting in greater variance but unbiased population

estimates (Sudman et al., 1996). The findings from Mathiowetz and Duncan (1986) illustrate the difference in reports based on estimation strategies as compared to episodic enumeration. In their study, population estimates of annual hours of unemployment for a 2-year reference period based on respondents' reports of unemployment hours were reasonably accurate. In contrast, when respondents had to report the months and years of individual spells of unemployment (requiring episodic enumeration) more than 60 percent of the individual spells of unemployment were not reported.

Several empirical investigations have identified means by which to improve the reporting of retrospective information for both episodic enumeration and estimation-based reports. These questionnaire design approaches include:

Event History Calendar. Work in the field of cognitive psychology has provided insight into the structure of autobiographic information in memory. The research indicates that "certain types of autobiographical memories are thematically and temporally structured within an hierarchical ordering" (Belli, 1998). Event history calendars have been found to be effective in reducing response error related to the reporting of what, when, and how often events occurred (Freedman et al., 1988). Whereas traditional survey instruments ask for retrospective reports through a set of discrete questions (e.g., "In what month and year did you last receive welfare payments?"), thereby emphasizing the discrete nature of events, event history calendars emphasize the relationship between events within broad thematic areas or life domains (work, living arrangements, marital status, child bearing and rearing). Major transitions within these domains such as getting married or divorced, giving birth to a child, moving into a new house, or starting a job, are identified by the respondent and recorded in such ways as to facilitate "an extensive use of autobiographical memory networks and multiple paths of memory associated with top-down, sequential, and parallel retrieval strategies" (Belli, 1998). If the question items of interest require the dating of several types of events, the literature suggests that the use of event history calendars will lead to improved reporting. For example, event history calendars could prove to be beneficial in eliciting accurate responses to questions such as "What was the year and month that you first received welfare cash assistance as an adult?"

Landmark Events. The use of an event history calendar is most beneficial if the questionnaire focuses on the dating and sequencing of events and behaviors across several life domains. In some cases, the questionnaire contains a limited number of questions for which the respondent must provide a date or a correct sequence of events. In these cases, studies have indicated that the use of landmark dates can improve the quality of reporting by respondents (Loftus and Marburger, 1983). Landmark events are defined as either public or personal landmarks; for some of these, the respondent can provide an accurate date (personal landmark

such as birthday, anniversary) whereas public landmarks can be dated accurately by the researcher. Landmarks are effective for three reasons: (1) landmark dates make effective use of the cluster organization of memory; (2) landmark dates may convert a difficult absolute judgment of recency to an easier relative judgment; and (3) landmark dates may suggest to the respondent the need to pay attention to exact dates and not simply imprecise dates. One way to operationalize landmark dates is to begin the interview with the respondent noting personal and/or public landmark dates on a calendar that can be used for reference throughout the interview.

Use of Records. If the information has not been encoded in memory, the response quality will be poor no matter how well the questions have been constructed. For some information, the most efficient and effective means by which to improve the quality of the reported data is to have respondents access records. Several studies report an improvement in the quality of asset and income information when respondents used records (e.g., Maynes, 1968; Grondin and Michaud, 1994; Moore et al., 1996). Two factors often hinder questionnaire designers from requesting that respondents use records: interviewers' reluctance and mode of data collection. Although in some cases interviewers have been observed discouraging record use (Marquis and Moore, 1990), studies that request detailed income and expenditure information such as the SIPP and the National Medical Expenditure Survey, have both reported success in encouraging respondents to use records (Moore et al., 1996). Record use by respondents is directly related to the extent to which interviewers have been trained to encourage their use by respondents. For telephone interviews, the fear is that encouraging record use may encourage nonresponse; a small body of empirical literature does not support this notion (Grondin and Michaud, 1994). One form of record to consider is the prospective creation of a diary that is referenced by the respondent during a retrospective interview.

Recall versus Recognition. Any free-recall task, such as the enumeration of all sources of income, is a cognitively more difficult task than the task of recognition, such as, asking the respondent to indicate which of a list of income sources is applicable to his or her situation. Consider the two approaches taken in examples 1 and 2:

> Example 1: In (PRIOR MONTH), did you receive any money or income from any other source? This might include (READ SLOWLY) unemployment insurance, workers' compensation, alimony, rent from a tenant or boarder, an income tax refund, foster child payments, stipends from training programs, grandparents' Social Security income, and so on.

Example 2: Next, I will read a list of benefit programs and types of support and I'd like you to tell me whether you or someone in your home gets this.

Food stamps
Medicaid
Child-care assistance
Child support from a child's parent
Social Security

In the first example, the respondent must process all of the items together; most likely after the first or second item on the list was read, the respondent failed to hear or process the remaining items on the list. Hence the list does not provide an effective recognition mechanism. In the second example, the respondent is given time to process each item on the list individually (the entire list consists of 20 items).

Complex Behavioral Experience. Simple behavioral experiences are relatively easy to report even over long reference periods whereas complex behavioral experiences can be quite difficult to reconstruct. For example, the experience of receiving welfare benefits continuously over a 12-month period is quite different from the experience of receiving benefits for 8 of the 12 months. The use of filter questions to identify those for whom the behavioral experience is complex would permit the questionnaire designer to concentrate design efforts on those respondents for whom the task is most difficult. Those with complex behavioral experiences could be questioned using an event history calendar whereas those for whom the recent past represents a steady state could be asked a limited number of discrete questions.

Recall Strategies. When respondents are asked to report a frequency or number of times an event or a behavior occurred, they draw on different response strategies to formulate a response. The choice of response strategy is determined, in part, by the actual number or frequency as well as the regularity of the behavior. Rare or infrequent events often are retrieved through episodic enumeration in which the respondent attempts to retrieve each occurrence of the event. Such strategies are subject to errors of omission as well as misdating of the event by the respondent. When the event or behavior of interest occurs frequently, respondents often will use some form of estimation strategy to formulate a response. These strategies include rule-based estimation (recall a rate and apply to timeframe of interest), automatic estimation (drawn from a sense of relative or absolute frequency), decomposition (estimate the parts and sum), normative expectations, or some form of heuristic, such as availability heuristic (based on the speed of retrieval). All estimation approaches are subject to error, but a well-designed questionnaire can both suggest the strategy for the respondent to use and attempt

to correct for the expected biases. For example, if the behavior or event of interest is expected to occur on a regular basis, a question that directs the respondent to retrieve the rule, and apply the rule to the time frame of interest, and then probes to elicit exceptions to the rule may be a good strategy for eliciting a numeric response.

Current versus Retrospective Reports. Current status most often is easier to report, with respect to cognitive difficulty, than retrospective status, so it is often useful to consider beginning questions concerning current status. Information retrieved as part of the reporting of current status also will facilitate retrieval of retrospective information.

REPAIRS FOCUSING ON PROBLEMS RELATED TO SOCIAL DESIRABILITY

Questions for which the source of the measurement error is related to perceived sensitivity of the items or the socially undesirable nature of the response often call for the use of question items or questionnaire modes that provide the respondent within greater sense of confidentiality or even anonymity as a means for improving response quality. The questionnaire designer must gauge the level of sensitivity or threat (or elicit information on sensitivity or threat through developmental interviews or focus groups) and respond with the appropriate level of questionnaire modifications. The discussion that follows attempts to provide approaches for questions of varying degrees of sensitivity, moving from slightly sensitive to extremely sensitive or illegal behaviors.

Reducing Threat Through Question Wording

Sudman and Bradburn (1982) provide a checklist of question approaches to minimize threat from sensitive questions. Among the suggestions made by the authors are the use of open questions as opposed to closed questions (so as to not reveal extreme response categories), the use of longer questions so as to provide context and indicate that the subject is not taboo, the use of alternative terminology (e.g., street language for illicit drugs), and embedding the topic in a list of more threatening topics to reduce perceived threat, because threat or sensitivity is determined in part by the context.

Alternative Modes of Data Collection

For sensitive questions, one of the most consistent findings from the experimental literature indicates that the use of self-administered questionnaires results in higher reports of threatening behavior. For example, in studies of illicit drug use, the increase in reports of use was directly related to the perceived level of

sensitivity, greatest for the reporting of recent cocaine use, less profound but still significant with respect to marijuana and alcohol use. Alternative modes could involve the administration of the questions by an interviewer, with the respondent completing the response categories using paper and pencil, or administration of the questionnaire through a portable cassette and self-recording of responses. More recently, face-to-face data collection efforts have experimented with CASID in which the respondent reads the questions from the computer screen and directly enters the responses and ACASI, in which the questions can be heard over headphones as well as read by the respondent. The latter has the benefit of not requiring the respondent to be literate; furthermore, it can be programmed to permit efficient multilingual administration without requiring multilingual survey interviewers. In addition, both computer-assisted approaches offer the advantage that complicated skip patterns, not possible with paper and pencil self-administered questionnaires, can be incorporated into the questionnaire. Similar methods are possible in telephone surveys, with the use of push-button or voice recognition technology for the self-administered portion of the questionnaire.

Randomized Response and Item Count Techniques

Two techniques described in the literature provide researchers with a means of obtaining a population estimate of an event or a behavior but not information that can be associated with the individual. Both were designed initially for use in face-to-face surveys; it is feasible to administer an item count approach in a telephone or self-administered questionnaire. The randomized response technique is one in which two questions are presented to the respondent, each with the same response categories, usually yes and no. One question is the question of interest; the other is a question for which the distribution of the responses for the population is known. Each question is associated with a different color. A randomized device, such as a box containing beads of different colors, indicates to the respondent which of the questions to answer, for which he or she simply states to the interviewer either "yes" or "no." The probability of selecting the red bead as opposed to the blue bead is known to the researcher. An example is as follows: A box contains 100 beads, 70 percent of which are red, 30 percent of which are blue. When shaken, the box will present to the respondent one bead (only seen by the respondent). Depending on the color, the respondent will answer one of the following questions: (Red question) Have you ever had an abortion? and (Blue question) Is your birthday in June? In a survey of 1,000 individuals, the expected number of persons answering "yes" to the question about the month of the birthday is approximately 1,000(.30)/12 or 25 persons (assuming birthdays are equally distributed over the 12 months of the year). If 200 persons said "yes" in response to answering either the red or blue questions, then 175 answered yes in response to the abortion item, yielding a population estimate of the percent of women having had an abortion as 175/(1000*.70) or 25 percent.

The item count method is somewhat easier to administer than the randomized response technique. In the item count method, two nearly identical lists of behaviors are developed; in one list *k* behaviors are listed and in the other list, *k +1* items are listed, where the additional item is the behavior of interest. Half of the respondents are administered the list with *k* items and the other half are offered the list with the *k +1* behaviors. Respondents are asked to simply provide the number of behaviors in which they have engaged (without indicating the specific behaviors). The difference in the number of behaviors between the two lists provides the estimate of the behavior of interest.

The major disadvantage of either the randomized response technique or item count method is that one cannot relate individual characteristics of the respondents with the behavior of interest; rather one is limited to a population estimate.

CONCLUSIONS

The empirical literature addressing response errors specifically among the low-income or welfare population is limited. However, if we couple those limited findings with results based on studies of the general population, some principles of questionnaire design to minimize response error emerge. At the risk of appearing to provide simple solutions to complex problems, we speculate on some guidelines to assist in the construction of questionnaires targeted at the low-income or welfare populations.

• Complex versus simple behavioral experience. One finding that is consistent throughout the literature indicates that complex behavioral experiences are more difficult to retrieve and report accurately than simple behavioral experiences. Despite this, questionnaire designers tend to treat all potential respondents the same, opting for a single set of questions for many questions, such as a single question or set of questions concerning annual earnings or amount of program support. One means by which to attempt to improve the reporting for those persons for whom the task is most difficult is to adopt, as suggested by Schaeffer (1994), the use of filter questions to determine the complexity of the experience, offering different follow-up questions for those with simple and complex behavior. For example, the person who has been employed continuously at a single job or unemployed continuously during a particular reference period easily can be identified and directed toward a different set of questions concerning earnings than the individual who has held several jobs, either concurrently or sequentially. Similarly, one can ask the respondent whether the amount of income from a particular income support program varies from month to month, with follow-up questions based on the response. Although this approach to questionnaire design deviates from the desire to "standardized" the measurement process, it acknowledges the need to be flexible within a standardized measurement process so as to maximize the quality of the final product.

• Simple, single-focus items often are more effective than complex, compound items. Whenever possible, a question should attempt to address a single concept. Questions that include the use of "and" or "or" or that end with exclusion or inclusion clauses often can be confusing to respondents. Although these questions often are constructed so as to minimize the number of questions read to the respondent (and therefore minimize administration time), we speculate that the use of several shorter questions is more effective, both from the perspective of administration time as well as the quality of the data. As an example, let's return to an earlier example:

> Since your welfare benefits ended in (FINAL BENEFIT MONTH), did you take part for at least one month in any Adult Basic Education (ABE) classes for improving your basic reading and math skills, or GED classes to help you prepare for the GED test, or classes to prepare for a regular high school diploma?

One means to improve this item would be as follows:

> Since (FINAL BENEFIT MONTH) have you taken any of the following classes?
> a. An Adult Basic Education class for improving basic reading and math skills? YES/NO
> b. A GED class to prepare for the GED test? YES/NO
> c. A class or classes to prepare for a regular high school diploma? YES/NO

If the "one month" qualifier offered in the original question was important analytically, each "yes" response could be followed up with a probe directed at the length of the class.

• Reduce cognitive burden whenever possible. Regardless of the population of interest, we know that, from a cognitive perspective, some tasks are easier to perform than others. Several means by which this can be accomplished include:

• Phrase tasks in the form of recognition rather than free recall. For example, asking the respondent to answer the question "Did you receive income from any of the following sources?" followed by a list of income sources is easier than asking the respondent to identify all income sources for the reference period of interest. Note that in asking a recognition question such as the one described, the ideal format would be to have the respondent respond "yes/no" to each income source, so only one item needs to be processed.

• Request information that requires estimation rather than episodic recall. For example, asking for the total number of jobs held during the reference period of interest requires less cognitive effort than asking for the starting and ending date of each job. If the latter information is needed to address analytic needs,

preceding the request with an estimation question may aid the respondent's retrieval of individual episodes.

- Request information in the format or metric used by the respondent. For example, earning information may be best reported when the most salient or most rehearsed metric is used by the respondent. For example, the findings by Borus (1970) and Smith (1997) that indicated a single broad-based question yielded a more accurate reporting by low-income respondents than a series of questions that required event-history type reconstruction of earnings simply may indicate that annual earnings are well rehearsed and more easily accessible to respondents than earnings related to any one job. One means by which to determine whether to ask the respondent about annual earnings, monthly earnings, or hourly earnings is to ask the respondent how he or she is best able to respond. Once again, this implies that tailoring the questionnaire to the respondent's circumstances may result in higher quality data.
- Focus on reference periods that are salient to the respondent. The 6-month period prior to exiting welfare may not necessarily be a particularly salient reference period, even though the date of termination of benefits may be quite salient. For reference periods that may not be salient to the respondent, the use of calendars or other records coupled with the identification of landmark events within the reference period may aid retrieval of information and the dating of events and behaviors.
- Provide the respondent with assistance in how to perform the task. For the most part, respondents rarely perform the task we are asking them to tackle. Instructions and feedback throughout the process can clarify the task for the respondent as well as provide feedback for appropriate respondent behavior. Instructions indicating that the questionnaire designer is interested in all spells of unemployment, including short spells lasting less than a week, provides an instruction to the respondent as well as additional time for the respondent to search his or her memory. Should the respondent provide such information, appropriate feedback would indicate that such detailed information is important to the study. Other forms of instruction could focus the respondent on the use of a calendar or other types of records.

In addition, we know from the literature that use of additional probes or cues stimulates the reporting of additional information. When there is interest in eliciting information from the respondent concerning short spells of employment or unemployment or odd or sporadic sources of income, repeated retrieval attempts by the respondent in response to repeated questions may be the most effective approach.

In some cases, the provision of some information may be preferable to no information from the respondent. Consider the case in which the respondent reports "don't know" in response to a question concerning earnings. One approach that has been effective is the use of broad-based followup questions in response to "don't know" items, for example, asking the respondent if his or her

earnings were more than or less than a specific amount, with subsequent followup items until the respondent can no longer make a distinction (see Hurd and Rodgers, 1998).

• Comprehension. The concepts of interest for many surveys of the low-income and welfare populations are fairly complex, for example, distinguishing among the various income support programs or determining whether sporadic odd jobs count as being employed. As indicated in several of the studies reviewed, research directed toward improving the comprehension of survey questions is greatly needed. For those developing questionnaires, this implies the need for iterative testing and pretesting, focusing on the interpretation of questions among members of the population of interest.

The empirical literature provides evidence of both reasonably accurate reporting of earnings, other sources of income, and employment as well as extremely poor reporting of these characteristics on the part of household respondents. The magnitude of measurement error in these reports is in part a function of the task as framed by the question. Careful questionnaire construction and thorough testing of questions and questionnaires can effectively identify question problems and reduce sources of error.

REFERENCES

Aquilino, W.
1994 Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly* 58:210-240.

Aquilino, W., and L. LoSciuto
1990 Effect of interview mode on self-reported drug use. *Public Opinion Quarterly* 54:362-395.

Barron, J., M. Berger, and D. Black
1997 *On the Job Training*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Belli, R.
1998 The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory* 6(4):383-406.

Belson, T.
1981 *The Design and Understanding of Survey Questions*. Aldershot, Eng.: Gower Publishing Company.

Bogen, K.
1995 Results of the Third Round of SIPP Cognitive Interviews. Unpublished manuscript, U.S. Bureau of the Census.

Borus, M.
1966 Response error in survey reports of earnings information. *Journal of the American Statistical Association* 61:729-738.
1970 Response error and questioning technique in surveys of earnings information. *Journal of the American Statistical Association* 65:566-575.

Bound, J., and A. Krueger
1991 The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics* 9:1-24.

Bradburn, N.
1983 Response effects. In *Handbook of Survey Research*, P. Rossi, J. Wright, and A. Anderson, eds. New York: Academic Press.

Burton, S., and E. Blair
1991 Task conditions, response formation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly* 55:50-79.

Cannell, C., G. Fisher, and T. Bakker
1965 Reporting of hospitalization in the health interview survey. *Vital and Health Statistics, Series 2, No. 6*. Washington, DC: U.S. Public Health Service.

Cannell, C., K. Marquis, and A. Laurent
1977 A summary of studies of interviewing methodology. *Vital and Health Statistics, Series 2, No. 69*. Washington, DC: U.S. Public Health Service.

Cannell, C., P. Miller, and L. Oksenberg
1981 Research on interviewing techniques. In *Sociological Methodology*, S. Leinhardt, ed. San Francisco: Jossey-Bass.

Carstensen, L., and H. Woltman
1979 Comparing earnings data from the CPS and employers' records. In *Proceedings of the Section on Social Statistics*. Alexandria, VA: American Statistical Association.

Cash, W., and A. Moss
1969 Optimum recall period for reporting persons injured in motor vehicle accidents. In *Vital and Health Statistics, Series 2, No. 50*. Washington, DC: U.S. Department of Health and Human Services.

Crowder, R.
1976 *Principles of Learning and Memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cynamon, M., and D. Camburn
1992 Employing a New Technique to Ask Questions on Sensitive Topics. Unpublished paper presented at the annual meeting of the National Field Directors Conference, St. Petersburg, FL, May, 1992.

David, M.
1962 The validity of income reported by a sample of families who received welfare assistance during 1959. *Journal of the American Statistical Association* 57:680-685.

Dibbs, R., A. Hale, R. Loverock, and S. Michaud
1995 *Some Effects of Computer Assisted Interviewing on the Data Quality of the Survey of Labour and Income Dynamics*. SLID Research Paper Series, No. 95-07. Ottawa: Statistics Canada.

Dodge, R.
1970 Victim Recall Pretest. Unpublished memorandum, U.S. Bureau of the Census, Washington, DC. [Cited in R. Groves (1989).]

Dreher, G.
1977 Nonrespondent characteristics and respondent accuracy in salary research. *Journal of Applied Psychology* 62:773-776.

Duncan, G., and D. Hill
1985 An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics* 3:508-532.

Forsyth, B., and J. Lessler
1991 Cognitive laboratory methods: A taxonomy. In *Measurement Error in Surveys*, P. Biemer, S. Sudman, and R.M. Groves, eds. New York: John Wiley and Sons.

Freedman, D., A. Thornton, D. Camburn, D. Alwin, and L. Young-DeMarco
1988 The life history calendar: A technique for collecting retrospective data. In *Sociological Methodology*, C. Clogg, ed. San Francisco: Jossey-Bass.

Gems, B., D. Gosh, and R. Hitlin
1982 A recall experiment: Impact of time on recall of recreational fishing trips. In *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Goodreau, K., H. Oberheu, and D. Vaughan
1984 An assessment of the quality of survey reports of income from the Aid to Families with Dependent Children (AFDC) program. *Journal of Business and Economic Statistics* 2:179-186.

Grondin, C., and S. Michaud
1994 Data quality of income data using computer assisted interview: The experience of the Canadian Survey of Labour and Income Dynamics. In *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

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

Halsey, H.
1978 Validating income data: lessons from the Seattle and Denver income maintenance experiment. In *Proceedings of the Survey of Income and Program Participation Workshop-Survey Research Issues in Income Measurement: Field Techniques, Questionnaire Design and Income Validation*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Hansen, M., W. Hurwitz, and M. Bershad
1961 Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute* 38:359-374.

Hardin, E., and G. Hershey
1960 Accuracy of employee reports on changes in pay. *Journal of Applied Psychology* 44:269-275.

Hill, D.
1987 Response errors around the seam: Analysis of change in a panel with overlapping reference periods. Pp. 210-215 in *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Hoaglin, D.
1978 Household income and income reporting error in the housing allowance demand experiment. In *Proceedings of the Survey of Income and Program Participation Workshop-Survey Research Issues in Income Measurement: Field Techniques, Questionnaire Design and Income Validation*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Horvath, F.
1982 Forgotten unemployment: recall bias in retrospective data. *Monthly Labor Review* 105:40-43.

Hurd, M., and W. Rodgers
1998 The Effects of Bracketing and Anchoring on Measurement in the Health and Retirement Survey. Institute for Social Research, University of Michigan, Ann Arbor, MI.

Jones, E., and J. Forrest
1992 Underreporting of abortions in surveys of U.S. women: 1976 to 1988. *Demography* 29:113-126.

Keating, E., D. Paterson, and C. Stone
1950 Validity of work histories obtained by interview. *Journal of Applied Psychology* 34:6-11.

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

Levine, P.
1993 CPS contemporaneous and retrospective unemployment compared. *Monthly Labor Review* 116:33-39.
Livingston, R.
1969 Evaluation of the reporting of public assistance income in the Special Census of Dane County, Wisconsin: May 15, 1968. In *Proceedings of the Ninth Workshop on Public Welfare Research and Statistics*.
Loftus, E., and W. Marburger
1983 Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition* 11:114-120.
London, K., and L. Williams
1990 A Comparison of Abortion Underreporting in an In-Person Interview and Self-Administered Question. Unpublished paper presented at the Annual Meeting of the Population Association of America, Toronto, April.
Lyberg, L., and D. Kasprzyk
1991 Data collection methods and measurement error: An Overview. In *Measurement Error in Surveys*, P. Biemer, S. Sudman, and R.M. Groves, eds. New York: Wiley and Sons.
Marquis, K., and J. Moore
1990 Measurement errors in SIPP program reports. In *Proceedings of the Annual Research Conference*. Washington, DC: U.S. Bureau of the Census.
Mathiowetz, N.
1986 The problem of omissions and telescoping error: New evidence from a study of unemployment. In *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.
Mathiowetz, N., and G. Duncan
1988 Out of work, out of mind: Response error in retrospective reports of unemployment. *Journal of Business and Economic Statistics* 6:221-229.
Maynes, E.
1968 Minimizing response errors in financial data: The possibilities. *Journal of the American Statistical Association* 63:214-227.
Mellow, W., and H. Sider
1983 Accuracy of response in labor market surveys: Evidence and implications. *Journal of Labor Economics* 1:331-344.
Menon, G.
1993 The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research* 20:431-440.
1994 Judgments of behavioral frequencies: Memory search and retrieval strategies. In *Autobiographical Memory and the Validity of Retrospective Reports*, N. Schwarz and S. Sudman, eds. New York: Springer-Verlag.
Moore, J., K. Marquis, and K. Bogen
1996 The SIPP Cognitive Research Evaluation Experiment: Basic Results and Documentation. Unpublished report, U.S. Bureau of the Census, Washington, DC.
Moore, J., L. Stinson, and E. Welniak
1999 Income reporting in surveys: Cognitive issues and measurement error. In *Cognition and Survey Research*, M. Sirken, D.J. Herrmann, S. Schechter, and R. Tourangeau, eds. New York: Wiley and Sons.
Morganstern, R., and N. Bartlett
1974 The retrospective bias in unemployment reporting by sex, race, and age. *Journal of the American Statistical Association* 69:355-357.

Murphy, L., and C. Cowan
1976 Effects of bounding on telescoping in the national crime survey. In *Proceedings of the Social Statistics Section*. Alexandria, VA: American Statistical Association.
Neter, J., and J. Waksberg
1964 A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association* 59:18-55.
Oberheu, H., and M. Ono
1975 Findings from a pilot study of current and potential public assistance recipients included in the current population survey. In *Proceedings of the Social Statistics Section*. Alexandria, VA: American Statistical Association.
O'Reilly, J., M. Hubbard, J. Lessler, P. Biemer, and C. Turner
1994 Audio and video computer assisted self-interviewing: Preliminary tests of new technology for data collection. *Journal of Official Statistics* 10:197-214.
Poterba, J., and L. Summers
1984 Response variation in the CPS: Caveats for the unemployment analyst. *Monthly Labor Review* 107:37-42.
Presser, S., and J. Blair
1994 Survey pretesting: Do different methods produce different results? *Sociological Methodology*. San Francisco: Jossey-Bass.
Presser, S., and L. Stinson
1998 Data collection mode and social desirability bias in self-reported religious attendance. *American Sociological Review* 63:137-145.
Rodgers, W., C. Brown, and G. Duncan
1993 Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association* 88:1208-1218.
Schaeffer, N.
1994 Errors of experience: Response errors in reports about child support and their implications for questionnaire design. In *Autobiographical Memory and the Validity of Retrospective Reports*, N. Schwarz and S. Sudman, eds. New York: Springer-Verlag.
Smith, J.
1997 Measuring Earning Levels Among the Poor: Evidence from Two Samples of JTPA Eligibles. Unpublished manuscript, University of Western Ontario.
Stinson, L.
1997 The Subjective Assessment of Income and Expenses: Cognitive Test Results. Unpublished manuscript, U.S. Bureau of Labor Statistics, Washington, DC.
Sudman, S., and N. Bradburn
1973 Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association* 68:805-815.
1982 *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass.
Sudman, S., N. Bradburn, and N. Schwarz
1996 *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.
Tourangeau, R., and T. Smith
1996 Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 60:275-304.
Tourangeau, R., L. Rips, and K. Rasinski
2000 *The Psychology of Survey Response*. Cambridge, Eng.: Cambridge University Press.

U.S. Bureau of the Census
1979 Vocational school experience: October, 1976. In *Current Population Reports Series P-70, No. 343*. Washington, DC: Department of Commerce.

Warner, K.
1978 Possible increases in the underreporting of cigarette consumption. *Journal of the American Statistical Association* 73:314-318.

Yen, W., and H. Nelson
1996 Testing the Validity of Public Assistance Surveys with Administrative Records: A Validation Study of Welfare Survey Data. Unpublished paper presented at the Annual Conference of the American Association for Public Opinion Research, May.