ON THE PATH TO QUALITY IMPROVEMENT IN SOCIAL MEASUREMENT: DEVELOPING INDICATORS OF SURVEY ERRORS AND SURVEY COSTS

Robert M. Groves, The University of Michigan and U.S. Census Bureau
Washington, D.C. 20233

The conflict between descriptive and analytic uses of survey and other observational data has stimulated a large literature in statistics and the social sciences (e.g., Deming, 1943; Brewer and Mellor, 1973; Hansen, Madow, and Tepping, 1983). Differences between them are most evident in the conditioning of inference on explicitly specified models, with analytic uses of surveys, and the conditioning of inference on various data collection design features, with descriptive uses of surveys. Following Groves (1989) let us refer to those with descriptive goals as **describers** and those with analytic goals as **modelers.**

This paper contends that, although historically the describers and modelers tended to use different survey designs and analytic plans, recent developments suggest that the two perspectives can be combined in useful ways to improve both the quality of the information from surveys and knowledge about the levels of quality. These developments include statistical and econometric models and insights from cognitive and social psychology. The statistical and econometric models are used to specify the nature of error properties of survey design features and to provide estimates of those errors for the survey analyst. The cognitive and social psychological insights are used to motivate design features whose sole purpose is to provide parameter estimates for those error models and in a longer run goal, guide efforts to reduce the errors.

This paper further acknowledges that design decisions about error measurement and error reduction must be made within constraints of resources available for the data collection effort. These constraints force evaluation of likely relative magnitudes of error components, prior to the design. These estimates should be used to guide allocation of resources to error measurement and error reduction efforts.

This paper is divided into 3 parts. Section 1 provides a review of optimization procedures of survey design within cost constraints for sampling error. A brief review of classical survey error structures is then presented. Section 2 enumerates the available indicators for the various errors reviewed in Section 1 and discusses the cost implications of using those indicators in surveys. Section 3 evaluates the practicality of a cost-error modeling perspective for design decisions involving data quality indicators.

1. Optimization of Sample Design Within Cost Constraints

Introductory sampling texts provide a simple design problem solved through the use of cost and error models. The designer is given a certain amount of money, C, to do the survey. Further, the strata for the sample have already been identified, and the decision has been made to draw simple random samples in each stratum. The question is what sample size should be drawn from each stratum in order to minimize the sampling variance of the sample mean, given the fixed cost.

The sampling variance of the sample mean in a stratified random design has a simple expression, a function of sample sizes in each stratum (say, n_h in the h-th stratum). The allocation of the sample cannot be allowed to cost more than C, and thus costs become a constraint within which solutions must be identified. To solve the problem a cost model needs to be developed which contains terms that are also present in the error model. In other words, we need to determine the costs of each of the design units in the error model (the n_h's). As is typical with this approach, each of the units which acts to improve the quality of the survey statistics (in this case numbers of sample elements), also brings with it a cost. The larger the sample size in any one stratum, the fewer the resources available to the researcher for other strata. In this simple problem there is only one set of cost components that varies across sample designs. All other terms are fixed as part of the essential survey conditions.

The cost model is parameterized in terms that are shared by the error model, the n_h's. The cost model could have been presented as a function of supervisors and interviewers salaries, of materials costs, computer time, etc., but that would have failed to represent clearly the fact that some costs rise as sampling variance decreases. Those are the only part of the cost equation that will determine the optimal allocation, the others will determine the overall sample size that can be purchased, but not what proportion of the units should be allocated to each stratum.

This paper addresses, among other issues, whether a formal cost-error perspective can or should be the only justification for efforts of survey designers to implement features whose sole purpose is the measurement of survey quality. It argues that weak indicators of nonsampling errors may offer valuable information to the survey analyst, while failing to meet the criteria of permitting use in formal cost-error models like those for sample design. To communicate the error structure of interest, the next section reviews various error components.

2. Available Indicators of Nonsampling Errors

Not all nonsampling error sources yield themselves to measurement. Some of them have yielded themselves to reduction, as evidenced, for example, by a literature evaluating interviewing training regimens. Some of the nonsampling errors, however, do yield themselves to measurement. The error measurement requires of the researcher, however, either the introduction of new measures in the data collection or the alteration of the statistical design of the selection or allocation of the sample. This section reviews current and past work in methods to enrich the information about the error properties of survey statistics.

2.1 Measurement of Nonresponse Errors With Selection Bias Models

The concern among modelers with the nonobservation of certain types of persons was well justified by Tobin (1958) and more recently by Heckman (1979). They note that biased estimates of population parameters in structural models can result from the systematic omission from the data of observations whose distribution on the dependent variable is different from that of those measured in the study. In nonprobability samples, the systematic omission may arise from the selection algorithm itself. In probability samples, the omission can arise at the time of measurement through failure to contact sample persons or the refusal of the sample person to cooperate with the sample request.

Most nonresponse cases are omitted not solely because of their values on a dependent variable, but as a result of some other process. In this case, the process is that of the ability of interviewers to contact sample persons and their willingness to be interviewed once contacted. The "selection" process is that which identifies elements of the sample to be interviewed. Such cases are termed the result of "incidental selection" in the language of Heckman (1979).

In addition to selection bias model use, there are two other statistical designs which have been suggested for nonresponse error measurement. Two phase sampling (Deming, 1953; Hansen and Hurwitz, 1958) utilizes a two step procedure. First, a sample is drawn and interviews are obtained with as many cases as possible, using the specified design. After the survey period is completed the remaining nonrespondents to the survey are subsampled (the double sample), and more expensive (and ideally, completely successful) procedures are used to seek interviews from them. Since they form a probability sample of the full set of nonrespondents, they provide estimates of the characteristics of nonrespondents which can be used for adjustment of survey statistics.

The use of more expensive methods can also be used from the beginning of the survey period, given to a probability subsample of the full sample. Differences in estimates between the group with the higher cooperation and that with the lower cooperation rate can be used for nonresponse bias adjustments. This is one of the motivations for the dual frame, mixed mode designs described by Lepkowski and Groves (1986).

The econometric literature provides methods to adjust for this error of nonobservation, but it does not provide good specifications of the selection model, the model that describes the process by which persons do or do not provide interview data. Similarly, the statistical literature on two phase sampling and split sample experiments for nonresponse does not give guidance to desirable stratification variables for the subsamples. This requires some social science theory. Two different literatures provide the time use data which describe the patterns of persons being at home and engaged in activities easily interrupted by interview requests, and b) the social and cognitive psychological literature on compliance and persuasion. These two literatures correspond to the two major sources of nonresponse error: noncontacts and refusals.

There are two obvious influences on a survey noncontact rate: the number of interviewer calls on a sample house and the "at-home" pattern of the household. Survey researchers have used their own methodology to attempt to predict the likelihood of contact by studying the times at which interviewers contacted respondents. There are also several published studies that describe the times at which different kinds of households or persons are at home.

The Weeks **et al.** (1980), the Weber and Burt (1972) and the Vigderhous (1981) studies share the design that data from a sample survey are analyzed to determine when sample households were contacted. The Weeks **et al.** work is based on a national area probability sample of over 22,000 housing units and reflect the result of the first call on the sample unit. About four percent were not successfully contacted after repeated calls. The Current Population Survey (CPS) data are from a national probability sample of about 6,000 households entering the sample for the first time. The 1960 Census data come from a time study covering all but the most rural areas of the country, about 20 percent of the population. Collapsing over all days of the week, the proportion of households with someone 14 years or older at home declines over the 16 years represented by the data. In 1976 larger proportions of homes are unoccupied at virtually every hour of the day relative to the 1960 and 1971 data. The authors cite reasons such as increasing proportion of females in the labor force, more multi-car families, and smaller numbers of persons per household. Across all three data sets, however, the proportion at-home is highest after 5:00 PM (generally about 5 to 10 percentage points higher than other hours). Saturday is distinctive from the weekdays in that the evenings have proportions at home similar to morning and afternoon hours (.55 to .65 range generally). Sunday tends to have low proportions at home in the morning (below .50), even lower than those during the weekday, but high proportions (over .70) after 6:00 PM.

Another source of data is time use studies, that have respondents report on their activities on a previous day. These data are thus more independent of the time interviewers choose to call on sample households. For example, Hill (1985) uses an "index of wakeful occupancy", which is the proportion of survey respondents who reported being at home and not sleeping. These are based on reports of respondents about their activities in the 24 hour period prior to the survey interview. Time use data show that this proportion over the different hours of a day, with the characteristic double moded distribution, one mode representing the at-home activities prior to leaving the housing unit in the morning and the second, those at home upon returning. The index declines in the evening at 11:00 - 11:30 PM. In addition to knowing that sample persons are home, we learn whether they are awake.

Such data might be preferable to survey call result data because 1) they may avoid the bias of call data being based on decisions of interviewers on appropriate times to call (thus probably tending to overestimate at home proportions), and 2) they provide data on what persons are doing when they are at home.

With regard to refusals, both sociological and psychological commentary is available. The perspective taken in a small sociological literature on refusals is that occupational and social roles, the strength of social networks, concerns about privacy, the saliency of a survey topic to an individual, and the degree of symmetry in the exchange of benefits between respondents and researchers all combine to influence the decision to cooperate with a survey request. This work takes a very different viewpoint than that of the survey methodological literature. It focuses on the individual sample person and examines what influences might come to bear on a decision to participate in the survey. The literature ignores to a large degree the influences arising from decisions of the survey designer (e.g., more callbacks, incentives).

A theoretical structure frequently taken to interpret survey participation is that of social exchange (see for example, Dillman, 1978; Goyder, 1988). In this view sample persons are seen to decide to cooperate with a survey request based on judged costs and benefits to them of that behavior. The costs of the survey participation include the time lost to other activities, the loss of privacy or control over information about oneself, the potential of being asked to reveal embarrassing attributes of oneself, and the engaging in a interaction whose agenda is controlled by another. The benefits of the interview include the supplying information that might improve society, the provision of assistance to the interviewer herself/himself, the opportunity to discuss a topic of personal interest, or the pleasure of interaction with another person. Tests of exchange theoretic hypotheses regarding surveys are difficult because they are based on values of the respondents themselves, which are seen to vary over individuals. Nonrespondents should have different valuations of the interaction with the interviewers, but their reactions cannot be measured, by definition. Dillman (1978) and others interpret several design decisions as likely to be universally valued by the sample person -- a personalized advance letter, commemorative stamps on mailed questionnaire envelopes. Such features, the argument goes, activate the exchange influence and yield more cooperation from sample persons.

Goyder notes several features of the survey interview atypical of other exchange relationships. The exchange in a survey is brief (extended somewhat by an advance letter or repeated callbacks), and thus probably relies on an "existing, natural, pre-conditioning and normative structure" (p.176). There is typically no opportunity to incorporate the survey request into a larger set of interactions between the researcher and sample person. It does suggest, however, that attempts to do so might be interpreted by some as enriching the relationship. For example, Dillman argues that follow-up of nonrespondents to seek their cooperation can itself more powerfully evoke exchange obligations by rewarding the nonrespondent with more attention from the researcher (Dillman, 1978). Those followups may be effective for some reluctant respondents because they successfully communicate to the sample person the importance that their participation is given by the researcher.

Another unusual feature of the exchange relationship in a survey context is the imbalance of power between interviewer and respondent. The interviewer initiates the interaction always; the interviewer often clearly communicates that repeated efforts at contact will result from an initial polite refusal; the interviewer clearly intends to set the agenda for the interaction. This asymmetry may make attempts by the researcher to increase the benefits of cooperation patently manipulative (e.g., monetary incentives seen as a token gift to obtain information which is much more valuable). If this is true, sample persons may not interpret the "favors", "gifts", and "acts of kindness" on the part of the interviewer as genuine. The perceived intent of the actions may be to increase cooperation. The exchange principle might then not be invoked.

A constituent concept in the exchange theory sketched above is the value placed on privacy by the sample person. Privacy is defined as "the right of an individual to keep information about herself or himself from others" (National Academy of Sciences, 1979, p.1). In order to make a decision about survey participation, each sample person must balance their right to privacy against the benefits of providing the desired information to the interviewer. From a content analysis of letters to the editors of a sample of newspapers in Canada and the United States, Goyder (1988) shows that issues of privacy are often mentioned in complaints against surveys and censuses. Without obtaining measures of nonrespondents' valuation of privacy rights, however, we cannot know how important they are as influences on cooperation.

Finally, many people believe that low response rates are caused by the society being "oversurveyed." They usually note that the populace has been saturated by survey attempts, and they cite market research surveys and political polls as chief offenders. Some acknowledge that their definition of surveys include telephone sales calls which begin with questions about the household.

In sum, sociological discussion of survey participation treats both possible effects of the frequency of social measurement on cooperation, of relative positions of interviewer and respondents in the social order, and of the value of privacy to the respondent. These in turn help shape reactions to the exchange relationship possible with the interviewer. Although this is the focus of the literature, there are auxiliary observations about the importance of the saliency of the topic to the sample person. This is a design feature chosen by the researcher, which also probably acts to influence the attraction of the exchange relationship.

The discussion above reviews various attributes of refusers to surveys, but it is unlikely that any of those attributes are the immediate **causes** of the refusal. Many of the sociological interpretations of survey refusals suggest that the person would be inclined toward certain psychological states which make cooperation less likely. Two literatures in social psychology study behaviors that resemble that of sample persons contacted by a survey interviewer. The first are studies of altruism and compliance. "Altruism" is operationalized as the provision of aid or assistance to another person without a formal request to do so. "Compliance" is the consent to a request for assistance by another. The second are studies of process of persuasion, how people respond to arguments for or against some belief or action on their part.

The Literature on Altruism and Compliance

The literature on compliance has explored a wide variety of behaviors from charitable contributions to consenting to donate bone marrow. Many of the experiments measure only the verbal consent to help the requestor, not the actual provision of the assistance. Many of the experiments use college students enrolled in psychology courses as subjects, with no concern about possible variation across subgroups in helping behavior.

This problem of generalizing from experimental to natural conditions exists for both the literatures on altruism and compliance and that on persuasion. The context of persuasion experiments is often laboratory environments with the subject reading a text arguing a specific position on some issue. Thus, the survey researcher must make judgements about whether the concepts found influential to subjects' decisions in those settings have relevance to the survey setting.

One way to assist in identifying the most useful concepts is organize them into higher order concepts and then logically test their applicability in a wide variety of settings. The work of Tversky and Kahneman (1973) describes a set of cognitive procedures which allow humans to make quick decisions based on insufficient information. These procedures, relying on what they call "heuristics", many times serve their users well. As Cialdini (1984) notes many of the notions of heuristics apply to decisions to cooperate with a request from salesmen, advertisers, waiters, and others seeking some action from us. Cialdini organizes the influences on compliance into six different concepts:

1) **reciprocation**, the tendency to favor requests from those who have previously given something to you,
2) **commitment and consistency**, the tendency to behave in a similar way over situations which resemble one another,
3) **social proof** or behavioral norms, the tendency to behave in ways similar to those like us,
4) **liking**, the tendency to comply with requests from attractive requestors,
5) **authority**, the tendency to comply with requests endorsed or given by those in positions of legitimate power,
6) **scarcity**, the tendency for rare opportunities to be more highly valued.

These influences have a tendency to be overused by the requestee and abused by the requestor to achieve his/her ends. For example, with regard to social proof, consumers may incorrectly judge that a popular automobile is well-built merely because others have purchased it. The mistake is that popularity may not be based on quality but on price. Conversely, advertisers can attempt to evoke the use of social proof by staging testimonials by "average" people about the quality of the product. They hope to influence a judgement that many people have carefully evaluated the product and found it superior. Hence, the viewer can be spared the burden of a detailed evaluation.

Using Social Science Concepts in Selection Models

How could these measures be collected, in order to specify the models for selection bias adjustment? For those measures on noncontact the call record data (times of calling on the sample unit) can be kept for both respondents and nonrespondent cases, and used as predictors of contact. Thus,

$$\mathbf{Y = XB + e}$$

where

X contains predictors of contact (e.g., number of calls on weekend mornings, number of calls on weekday evenings, etc.),

B contains parameters of the selection model,

Y is the likelihood of interview.

For measures of social and cognitive psychological influences interviewer observations and other measures may be required. On all respondent and nonrespondent cases, characteristics of the interviewer can be used as predictors, characteristics of the arguments given by the interviewer can be recorded, characteristics of the reaction of the sample person to the arguments, the presence of others during the description can be given, and other characteristics can be recorded. This requires an assembly of refusal correlates for respondent and nonrespondent cases. In essence, this is a supplementary questionnaire, used to collect measures whose sole purpose is error measurement and adjustment.

What are the cost implications of these attempts to measure nonresponse error? Most of the independent variables proposed above for selection models require observation or question-asking by the interviewer. This takes time. The number of minutes of respondent contact will increase. The amount of time spent in post contact recording of information may increase. The increase in time is applicable to both respondent and nonrespondent cases because predictor values are needed for both sets. One possible cost model is based on costs of individual indicators:

$C_{nr} = C_o + \text{SUM}(d_i + a_i)$

where

C_o are fixed costs of materials,

d_i is the cost of development the i-th nonresponse indicator,

a_i is the cost of interviewer time in observing or asking about the required indicator,

Optimization of the nonresponse indicator design might occur by maximizing the fit of the selection model, subject to the cost constraints of the cost model.

2.2 Asking Questions Whose Purpose is to Measure Comprehension of Substantive Questions

Recent work in laboratory testing of survey questionnaires has illustrated the utility of having the respondent reflect on the cognitive tasks demanded by the questions (see e.g., Royston et al., 1986). Some of this work is utilizing techniques of protocol analysis of memory retrieval experiments, having the respondent "think-aloud" to articulate the process by which an answer is determined to the question. Other attempts utilize probe questions which follow the original question. These sometimes are similar to the "random probe" suggest by Schuman (1966); other times they ask the respondent to report the perceived intent of the question, as they interpreted it. Finally, despite the demonstrated fallibility of human judgments about the accuracy of an answer in some circumstances (Tversky and Kahneman, 1974), there is work showing confidence ratings by the respondent on survey answers might be useful correlates of response error.

An early study by Ferber (1956) does provide an empirical link between attributed meaning of a question by a respondent and their answers. Ferber had interviewers ask attitudinal questions about specific political issues (e.g., "What is your attitude toward allowing labor to have a guaranteed annual wage? For, Against, Neutral, Don't Know."). These questions were then followed by a probe about the reason for the answer (i.e., "Why?"). The third question was one concerning the perceived meaning of the issue (e.g., "As you interpret it, what do the unions mean by a guaranteed annual wage?"), in the form of an open question. Judges examined the answers to the last question to code whether the respondent had a correct idea of the basic issue. Whether or not respondents knew the meaning of terms in the question, they were willing to provide the requested opinion. Large portions of those who later admit ignorance about the meaning of the issue provided an attitudinal response (from 14 percent to 83 percent across the four issues). Those who are misinformed (provide an incorrect definition) behave on the attitudinal question very similarly to those who know the meaning of the term.

Ferber's data also permit observations of whether the attitudes expressed by those who correctly interpreted the terms were different than others. Those using the intended meaning offer distinctive opinions relative to the full sample. For example, only a third of those who understand the meaning of "guaranteed annual wage" support it, but almost half (46.3 percent) of the total sample does. Ferber does not examine multivariate relationships involving these attitudinal variables, so we are not informed about the impacts of respondent ignorance on analytic statistics. The results offer strong support for Belson's hypotheses that the respondent will answer survey questions despite little understanding of the question.

Another example of using the respondents to obtain information about the meaning of words in questions is that described by Martin (1986). The National Crime Survey asks respondents to report criminal victimizations that occurred in the six months prior to the interview. This is communicated to the respondents at the beginning of the interview by

> "Now I'd like to ask some questions about crime. They refer only to the last 6 months -- between _____ 1, 19XX and _____, 19XX. During the last 6 months, did anyone break into ..."

After the victimization questions, several minutes of other questions were asked of the respondent. At the end of the interview, a set of debriefing questions were asked of the respondents (these are listed as Figure 4). One asked the respondent what they understood the reference period to be. Approximately 5 percent gave an answer different from the six month period; about 15 percent replied that they didn't know.

What is a desirable error model related to these efforts? One simple approach is to model the reported value to the question as

$$y_i = Y_i + e_i$$

where

e_i, the response error of respondent i, is itself a function of the followup question,

$$e_i = E + bx_i$$

where

x_i is the response to the followup question,

E is some base error rate in the population,

b is a parameter of the error model.

What are the cost implications of this attempt to measure errors associated with comprehension? These efforts increase the length of the interview. It is clear that the total effect of that is not merely the number of interviewer minutes required to administer the question. There are nonlinear effects of increasing questionnaire length that affect interviewer decisions about whether to attempt contact with a formally nonrespondent case or to wait for another day. Thus,

$$C_q = C_o + SUM(IM_i) + bMI$$

where

C_o are fixed costs related to requesting an interview of any length from a sample person,

I is the average per minute interviewer cost,

M_i are the number of minutes required to administer the i-th question,

M is the total length of the interview, including the error indicator questions,

b is a parameter reflecting the increase in minutes per interview as a function of total length.

The design optimization procedure might maximize differences in goodness of fit statistics for structural models because of the presence of the error indicators, subject to the cost constraint. This is clearly a different design decision rule than error minimization within cost constraints. It has less appeal under the conditions of an erroneous error indicator. It merely places a value on sensitivity coefficients and fit of a model to the presence of another predictor. If a plausible error model can be constructed using the indicator as a valid measure of error, then the design criterion is a good one.

2.3 Multiple Indicator Designs for Measuring Error

The use of multiple indicators of latent traits has a strong tradition in quantitative social science. It has also been used for direct estimates of correlated measurement errors, see Andrews (1984), Herzog and Andrews (1986). As with the other errors, randomization and split sample methods have also been used to address some error properties of questions (e.g., Schuman and Presser, 1981). The difference between these approaches concerns the ability to estimate covariances between the different indicators (and thus estimates of correlated response variance, given a particular measurement model). The split sample work has focused more clearly on the estimation of bias terms in measurement error, fixed weaknesses of survey questions which have consistent effects over replications.

The measurement error models for the two approaches are somewhat different. Those employing split sample methods generally fail to make their measurement models explicit, but appear to assert the following:

Response = True Value + Form Effect + Random Error

$y_{ij} = X_i + M_{ij} + e_{ij}$

where

y_{ij} = response obtained for the i-th person using the j-th method or form,

X_i = true value of the characteristic for the i-th person,

M_{ij} = effect on the response of the i-th person of using the j-th method,

e_{ij} = deviation for the i-th person from the average effect of the j-th method.

Those using a multiple indicator approach to measurement error, often use a multitrait multimethod procedure to provide estimates of measurement error variance. This views each response to a survey question as a function of the true value of the trait and a method effect:

Response = Population Mean + Influence of True Value + Method Effect + Random Error

$y_{ijkm} = m_k + b_{km}X_{ik} + a_{jm}M_{ij} + e_{im}$

where

y_{ijkm} = the observed value of the i-th person using the j-th method to measure the k-th characteristic using the m-th indicator,

m_k = the mean value of the k-th characteristic for the population studied,

b_{km} = "validity" coefficient for the m-th indicator of the k-th underlying trait,

X_{ik} = for the i-th person, the true value of the k-th characteristic,

a_{jm} = "method effect" coefficient for the m-th indicator of the j-th method,

M_{ij} = for the i-th person, the common effect of using the j-th method,

e_{im} = a random deviation for the i-th person on the m-th indicator.

What are the cost implications of these attempts to measure errors associated with question wording? A cost model that would apply to this case is

$C = C_o + (C_{oi} + C_{mi}m)n$

where

C_o = all costs that are not a function of the number of questions or number of respondents used,

C_{oi} = the cost of administering all the other questions in the questionnaire to a single respondent (i),

C_{mi} = the cost of administering each scale question to a single respondent,

m = mean number of items administered.

Design optimization in this case has been addressed previously in a simple case of an additive scale, by Lord and Novick (1968). The optimization procedure minimizes the variance of the scale value, given a cost constraint from the model above.

2.4 Interviewer Observations to Measure Respondent Behavior Correlated with Measurement Error

The final set of methods to get empirical estimates of survey error are observations taken by the interviewer about the behavior of the respondent. By far, the largest use of survey interviewers has been only to orally present survey questions and record responses. There has been little utilization of observation skills of survey interviewers to collect information additional to that of the survey questions. Some examples which do exist are the U.S. Consumer expenditure survey asking whether respondents consulted bills in answering questions about purchases of various sorts, the requirement that interviewers record probes that they found necessary to give or information they supplied in order for the respondent to answer the question (see SRC Interviewer's manual), and the use of interviewer judgments after the completion of the interview.

In a reverse record check survey of reporting of unemployment spells and salary and income data, the interviewers were asked at the end of the interview to answer questions about the respondent behavior during the interview. These questions concerned the entire interview, not a particular question and thus are a weak test of the ability of the interviewer to observe behavior correlated with response errors. Interviewers were asked to rate the respondents' understanding of the questions, their attempt to give accurate responses, their reluctance to begin the interview, and their asking about the likely length of the interview. These answers were used as predictors of differences between the respondent answers and record based data on the respondent. The attempt to be accurate (as perceived by the interviewer) was not found to be a useful predictor of response error for the prior year's income or that of 2 years ago, but was for years earlier than that. Something observable about the respondent behavior was related to the quality of their responses on such difficult to recall facts.

Whether interviewer observations can be useful indicators of measurement error in a wide variety of circumstances is largely unknown. It is the author's impression that little use is made of the observational powers of interviewers in most survey settings. Hence, it appears to be an area ripe for investigation.

The error and cost models applicable for these kinds of indicators are very similar to those for questioning of the respondent to obtain error indicators. The optimization through maximization of sensitivity of structural model estimates, given the specified measurement model, also seems appropriate.

3. Prospects of Minimizing Sampling and Nonsampling Error Subject to Fix Costs

The traditional sampling approach to design optimization uses an error model reflecting sampling variance and a cost model which is a function of terms in the error model. Minimization of sampling error within fixed cost is the typical goal of the exercise.

The error model can be stated because of the sampling theory which links design features to the magnitude of sampling error. That theory requires the existence of a sample design which provide also provides measures of the sampling error from each trial. That is, under probability sampling we can specify the relationship between stratification, clustering, and assignment of probabilities of selection to target population units, on one hand, and the resulting sampling variance of sample based statistics. It also is true that use probability samples provides estimates of the sampling variance when they are implemented.

The situation in the model based estimates of nonsampling error is inevitably somewhat different. At the present time, there is no well accepted model between any of the indicators of nonsampling error reviewed above and the magnitude of the error (e.g., if respondents in a debriefing interview are asked a question about the reference period used in the interview, we cannot specify the functional form of the relationship between a response to the question and measurement bias or variance). Further, for many of the indicators design features have not been identified to affect these error indicators (e.g., what interviewer behavior or question form is needed to improve respondent comprehension of the reference period).

If error indicators cannot be used in a straightforward way to reduce the error in question, why should they be collected? The first reason is that, if the error model is well specified, they will provide information cautioning the analyst about the limits of inference from the survey analysis. That is, the estimates of population parameters, both structural model parameters and descriptive finite population parameters, will have smaller bias or variance, in the presence of information provided by the error indicators. This alone should have merit.

The second reason follows the current practice in statistical quality control in manufacturing settings. There, imperfect indicators of product quality must often be used because perfect indicators require destructive tests. Monitoring of the values of these indicators occurs across time. Target values are set for the indicators. When these are repeatedly missed, intervention occurs. Because the error indicators are not directly linked to features which produce the errors, the intervention sometimes requires investment in "research" to identify what design features can be changed to reduce the error.

This situation is quite similar to that of the various indicators of nonsampling error reviewed above. Many of the indicators have surface appeal as reflecting breakdowns of the assumptions of the survey process (e.g., that questions mean the same to all respondents). The empirical relationship between values on the indicators and magnitudes of some survey bias or variance is not known, however. If values are large for these indicators, repeatedly over time for a particular survey design or a particular survey staff, then intervention is warranted. The intervention will require research into the sensitivity of the indicator to a particular error source and design features to reduce the error. From such research both better designs (with smaller errors) and better indicators of the errors may result.

References

Andrews, F.M., "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach," **Public Opinion Quarterly**, 48, 1984, pp. 409-442.

Andrews, F.M., "The Quality of Survey Data as Related to Age of Respondent," **Journal of the American Statistical Association**, June, 1986, 81, 394, pp. 403-410.

Cialdini, R.B., **Influence: How and Why People Agree to Things**, New York, William Morrow, 1984.

Dillman, D., **Mail and Telephone Surveys**, Wiley, 1978.

Ferber, Robert, "The Effect of Respondent Ignorance on Survey Results," **Journal of the American Statistical Association**, 1956, Vol. 51, No. 276, pp. 576-586.

Goyder, J., **The Silent Minority**, Boulder, CO: Westview Press, 1988.

Groves, R.M., **Survey Errors and Survey Costs**, Wiley, 1989.

Hansen, Morris H., William G. Madow, and Benjamin J. Tepping, "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," **Journal of the American Statistical Association**, December, 1983, Vol. 78, No. 384, pp. 776-793. And Comments following paper.

Heckman, J.J., "Sample Selection Bias as a Specification Error," **Econometrica**, 45, 1979, pp. 153-161.

Lepkowski, J.M., and R.M. Groves, "A Mean Square Error Model for Dual Frame, Mixed Mode Surveys," **Journal of the American Statistical Association**, 1986, 81, 396.

Lord, F. and Novick, **Statistical Theories of Mental Test Scores**, Chapters 1-3, Addison-Wesley, 1968.

Martin, E., **Final Report on the National Crime Survey Redesign**, U.S. Bureau of Justice Statistics, 1986.

Royston, P. et al., "Questionnaire Design Research Laboratory," **Proceedings of the Survey Research Methods Section of the American Statistical Association**, 1986.

Schuman, Howard, "The Random Probe: A Technique for Evaluating the Validity of Closed Questions," **American Sociological Review**, 1966, Vol. 31, pp. 218-222.

Schuman, H. and S. Presser, **Questions and Answers**, Academic Press, New York, 1981.

Tobin, J., "Estimation of Relationships for Limited Dependent Variables," **Econometrica**, 2b, 1958, pp. 24-36.

Tversky, A., and Kahneman, D., "Availability: A heuristic for judging Frequency and Probability," **Cognitive Psychology**, 1973, 5, pp. 207-232.

Tversky, A., and Kahneman, D., "Judgment Under Uncertainty: Heuristics and Biases, **Science**, 1974, 184, pp. 1124-1131.

Vigderhous, Gideon, "Scheduling Telephone Interviews: A Study of Seasonal Patterns," **Public Opinion Quarterly**, Vol. 45, Summer 1981, pp. 250-259.

Weber, D., and Burt, R.C., **Who's Home When**, U.S. Bureau of the Census, 1972.

Weeks, M. et al., "Optimal Times to Contact Sample Households", **Public Opinion Quarterly**, Spring, 1980, pp. 101-114.