TOTAL SURVEY ERROR: DESIGN, IMPLEMENTATION, AND EVALUATION

Author(s): PAUL P. BIEMER

Source: *The Public Opinion Quarterly*, 2010, Vol. 74, No. 5, Total Survey Error (2010), pp. 817-848

Published by: Oxford University Press on behalf of the American Association for Public Opinion Research

Stable URL: https://www.jstor.org/stable/40985407

Public Opinion Quarterly, Vol. 74, No. 5, 2010, pp. 817–848

TOTAL SURVEY ERROR
DESIGN, IMPLEMENTATION, AND EVALUATION

PAUL P. BIEMER*

Abstract The total survey error (TSE) paradigm provides a theoretical framework for optimizing surveys by maximizing data quality within budgetary constraints. In this article, the TSE paradigm is viewed as part of a much larger design strategy that seeks to optimize surveys by maximizing *total survey quality*; i.e., quality more broadly defined to include user-specified dimensions of quality. Survey methodology, viewed within this larger framework, alters our perspectives on the survey design, implementation, and evaluation. As an example, although a major objective of survey design is to maximize accuracy subject to costs and timeliness constraints, the survey budget must also accommodate additional objectives related to relevance, accessibility, interpretability, comparability, coherence, and completeness that are critical to a survey's "fitness for use." The article considers how the total survey quality approach can be extended beyond survey design to include survey implementation and evaluation. In doing so, the "fitness for use" perspective is shown to influence decisions regarding how to reduce survey error during design implementation and what sources of error should be evaluated in order to assess the survey quality today and to prepare for the surveys of the future.

Introduction

Total survey error (TSE) refers to the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data. In this context, a survey error is defined as the deviation of a survey response from its underlying true value. A related term—survey accuracy—is defined as the deviation of a survey estimate from its underlying true parameter value. Survey errors can

PAUL P. BIEMER is Distinguished Fellow at RTI International, Research Triangle Park, NC, USA. He is also Associate Director for Survey Research and Development in the Odum Institute for Research in Social Science, University of North Carolina, Chapel Hill, NC, USA. *Address correspondence to Paul Biemer, RTI International, P. O. Box 12194, Research Triangle Park, NC 27709-2194, USA; e-mail: ppb@rti.org.

doi: 10.1093/poq/nfq058

arise from the survey frame deficiencies, the sampling process, interviewing and interviewers, respondents, missing data, and coding, keying, and editing processes. Survey error is problematic because it diminishes the accuracy of inferences derived from the survey data. A survey estimator will be accurate if it has a small bias and variance, which occurs only if the influence of TSE on the estimate is small.

The *total survey error paradigm* (see, for example, Platek and Särndal 2001 and the ensuing discussions) refers to the concept of optimally allocating the available survey resources to minimize TSE for key estimates. Ideally, to apply the TSE paradigm, the major sources of error should be identified so that the survey resources can be allocated to reduce their errors to the extent possible, while still satisfying specified costs and timeliness objectives.

The TSE paradigm is part of the much broader concept of *total survey quality*, which considers the *fitness for use* of an estimate. The "fitness for use" concept (Juran and Gryna 1980) recognizes that producers and users of survey data often perceive survey quality from very different perspectives. Producers place a high priority on data quality (e.g., large sample size, high response rate, internally consistent responses, good coverage of the target population) and may allocate a large portion of the survey budget to achieve a high level of accuracy for some key estimates. Data users often take accuracy for granted and place a higher priority on attributes such as the timeliness, accessibility, and usability of the data, as well as questionnaire content that is highly relevant to their research objectives. These two perspectives suggest that survey quality is a complex, multidimensional concept that goes beyond TSE. Juran and Gryna (1980) identify two distinct facets of the general concept of quality: (a) freedom from deficiencies; and (b) responsiveness to customers' needs. For most surveys, (a) is consistent with the TSE paradigm; however, (b) can be achieved only by giving appropriate emphasis in the survey design to attributes that will result in high user satisfaction; in particular, data accessibility and clarity, timely delivery, and relevant data items that are comparable across repeated surveys and regions of the country, as well as across demographic groups and analytic domains.

Assigning lower priorities to the user dimensions of survey quality can result in data that are released behind schedule, difficult and costly to access, and inadequately documented. To the user, the data may be unfit for use. For example, for a continuing survey, changes in the methodology may produce data that are no longer comparable to earlier data releases, leaving the interpretation of time trends muddled as real changes are confounded by methodological artifacts. Or, important items on the questionnaire may be eliminated, thus weakening the relevance of the data to a substantial user group. The data may be accurate, but they lack *total survey quality*; that is, quality from both the producer and user perspectives. This situation is likely to result in users that are dissatisfied with the data products.

In the late 1970s, Lyberg et al. introduced quality dimensions that went beyond accuracy and were intended to embody the concept of fitness for use (Lyberg, Felme, and Olsson 1977). By the mid-1990s, some government statistical agencies began developing definitions for survey quality that explicitly take into account the multidimensionality of the concept (see, for example, Fellegi 1996). Such definitions are referred to as "survey quality frameworks." Today, most national statistical offices in Europe, as well as Eurostat, Australia, Canada, New Zealand, and the U.S. Census Bureau, are using very similar survey quality frameworks to some extent. Interestingly, nongovernmental survey organizations in both Europe and the United States have been slow to adopt the concept. The dimensions of a quality framework can vary (somewhat subtly in most cases) from organization to organization and can be a topic of considerable debate. Nevertheless, most frameworks contain a subset of the nine dimensions shown in table 1.

The next section describes some uses of the total survey quality framework, including a strategy for designing surveys that maximizes total survey quality. This is achieved by optimally balancing the dimensions of survey quality within the survey budget and schedule. Sections 3 and 4 describe the sources of error that reduce survey accuracy and how they can be summarized by the mean squared error. Section 5 discusses survey design principles within the TSE paradigm, Section 6 discusses the concept of process quality and its relationship to TSE, and Section 7 describes some options for assessing total survey quality. Finally, Section 8 concludes with a summary of the essential ideas.

Survey Design within the Total Survey Quality Framework

Survey organizations have used survey quality frameworks in various ways. Primarily it has been used as a checklist for the assessment of survey quality (i.e., to evaluate how well a data-collection program achieves the goals or

Table 1. Common Dimensions of a Survey Quality Framework

Dimension	Description
Accuracy	Total survey error is minimized
Credibility	Data are considered trustworthy by the survey community
Comparability	Demographic, spatial, and temporal comparisons are valid
Usability/Interpretability	Documentation is clear and metadata are well-managed
Relevance	Data satisfy users needs
Accessibility	Access to the data is user friendly
Timeliness/Punctuality	Data deliveries adhere to schedules
Completeness	Data are rich enough to satisfy the analysis objectives without undue burden on respondents
Coherence	Estimates from different sources can be reliably combined

requirements stated for each dimension). This implies that an evaluation should be conducted to collect data on quality indicators and metrics for each dimension. Some dimensions (such as accessibility) are qualitative and difficult to quantify, and thus a single metric summarizing the quality across all dimensions would be difficult to construct. Instead, quality reports or declarations have been developed that provide information on how well a survey satisfies specific goals for each dimension. The quality report might include a description of the strengths and weaknesses of a survey organized by dimension, with emphasis on sampling errors; nonsampling errors; key release dates for user data files; user satisfaction with data dissemination, availability, and contents of the documentation; and special features of the survey approach that may be of importance to most users. Extended versions of such reports, called *quality profiles*, have been produced for a number of surveys (see Doyle and Clark 2001 and Kasprzyk and Kalton 2001 for discussions of this approach).

Another important use of the quality framework is in the design of a survey. Ideally, the survey design should specify actionable and achievable objectives for each quality dimension, in accordance with both user and producer requirements. Budgets, personnel, and other resources can then be allocated to the various survey tasks and processes, as appropriate, to achieve these objectives. Thus, the optimal survey design is one that is best in terms of both user and producer quality dimensions—in other words, a design that optimizes total survey quality. In this way, the producer's goals of data accuracy and methodological credibility are explicitly and optimally balanced against the often competing goals for the other quality dimensions in table 1. Optimizing total survey quality requires that the quality goals for each dimension are clearly specified and the approaches of achieving these goals are optimally designed and budgeted.

One approach proposed by Biemer and Lyberg (2003) treats the user dimensions as constraints and maximizes data accuracy subject to those constraints. To illustrate, suppose that in addition to accuracy, the quality framework for a survey consists of three dimensions that have a substantial impact on costs (e.g., timeliness, accessibility, comparability). An optimal balance for a survey within this framework maximizes data accuracy, while ensuring that explicit objectives developed for the other three dimensions are accomplished within the survey budget. For example, the survey design may specify that data collection for the survey should be completed within nine months, and that data files will be released to the public within 15 months. The design may specify that data files will be provided for download online with full documentation at the time of release. Further, for the sake of comparability, methodologies used in previous implementations of the survey should be continued in the new implementation. The survey budget must take into account these objectives in the allocation of resources.

Let C_T be the total budget for the survey and C_U denote the combined, estimated costs for achieving the specified objectives for the user dimensions of

timeliness, accessibility, and comparability. The remaining budget (i.e., $C_A = C_T - C_U$) is the budget available to maximize accuracy. The task for the survey designer is to implement the data collection, data processing, weighting, and estimation phases of the survey to maximize accuracy, while ensuring that survey costs do not exceed C_A and the time from the start of data collection to the release of data files does not exceed 15 months. In this manner, the design specifications for data collection, data processing, weighting, and estimation minimize TSE subject to cost and timeliness (15 months) constraints. This approach attempts to maximize the total survey quality once the design objectives and specifications under each dimension are set in accordance with both user and producer requirements.

In actual practice, the total survey quality optimization strategy is iterative. For example, the designer may determine that the remaining budget, C_A, and/or schedule are inadequate for achieving an acceptable level of accuracy. Ideally, the survey sponsor would provide additional funding or allow more time to achieve a higher level of accuracy. But assuming the budget and schedule are fixed, the survey designer should revisit the objectives under the other quality dimensions to determine how resources might be reallocated in order to achieve greater accuracy. Of course, this should be done so that the impact on the most important user quality dimensions is minimized.

Sources of Error

As noted in the previous section, the goal of optimal survey design can be stated simply as minimizing TSE subject to costs and timeliness constraints that are consistent with other user-centric quality dimensions. Careful planning is required for allocating resources to the various stages of the survey process so that the major sources of error are controlled to acceptable levels. The goal is not to conduct every stage of the survey process as error-free as possible, because that would entail exceeding the survey budget and/or schedule by a considerable margin. Even under the best circumstances and given an unlimited budget and time, the potential for survey errors will always remain in some operations. Instead, the goal is to avoid the most egregious errors and control other errors to the extent that remaining errors are mostly inconsequential and tolerable.

As an example, more extensive interviewer training may be costly, but still necessary in order to minimize serious interviewing errors in a field survey. To afford these costs, quality control activities that would normally be in place to control data-processing and file-preparation errors may have to be reduced. Similarly, to afford the nonresponse follow-up activities required for reducing nonresponse bias, a reduction may be taken in the survey pretesting phase or in the length of the interview. These design choices and tradeoffs require an understanding of the major sources of survey error, their relative importance to

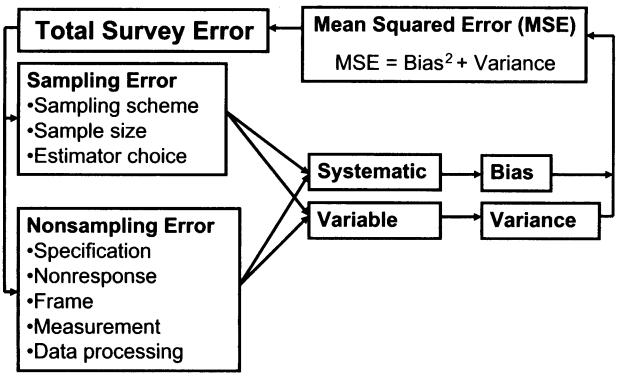


Figure 1. Total Survey Error, Its Components, and the Mean Squared Error.

data quality, how they can be controlled by optional features of the design, and the costs associated with these design features.

Addressing the most serious errors by judicious survey design is facilitated by decomposing the error to smaller and smaller components. One such decomposition, shown in figure 1, partitions the TSE first into sampling error and nonsampling error. Sampling error can be partitioned into error sources attributable to the sampling scheme (e.g., multistage or multiple-phase sample), sample size, and the choice of estimator (e.g., a ratio or regression estimator, levels of post-stratification). Nonsampling error can be further decomposed into specification error, frame error, nonresponse error, measurement error, and processing error. To allow for new error sources that may emerge as methods and technologies change, these five components can be broadly defined so that they encompass essentially all sources of nonsampling error in a survey. Further decomposition of both types of survey error is usually needed to better target specific sources of error (see Section 4). These error sources, considered in some detail in Biemer and Lyberg (2003) and Groves (1989), will be only briefly summarized here.

SPECIFICATION ERROR

A *specification error* arises when the concept implied by the survey question differs from the concept that should have been measured in the survey. When this occurs, the wrong construct is being measured and, consequently, the wrong parameter will be estimated by the survey, which could lead to invalid inferences. Specification error is often caused by poor communication between the researcher (or subject-matter expert) and the questionnaire designer.

An example of specification error is in the measurement of unemployment in the Current Population Survey (CPS) (U.S. Department of Labor, Bureau of

Labor Statistics, and U.S. Department of Commerce, Bureau of the Census 2002). For the Bureau of Labor Statistics (BLS), an important distinction among unemployed persons is whether they were "looking for work" or "on layoff." Persons on layoff are defined as those who are separated from a job and await a recall to return to that job. Persons who are "looking for work" are the unemployed who are not on layoff and who are pursuing certain specified activities to find employment. Prior to 1994, the CPS questionnaire did not consider or collect information on the expectation of recall from persons who indicated that they had been laid off. Rather, unemployed persons were simply asked, "Were you on layoff from a job?" This question was problematic because, to many people, the term "layoff" could mean permanent termination from the job rather than the temporary loss of work the BLS economists were trying to measure. (See Biemer 2004 for an extensive discussion and analysis of this problem.)

BLS redesigned this question in 1994 to clarify the concept of layoff. Currently, unemployed persons are asked, "Has your employer given you a date to return to work?" and "Could you have returned to work if you had been recalled?" These questions brought the concept of "on layoff" in line with the specification being used by BLS economists and produced slightly different estimates of unemployment.

MEASUREMENT ERROR

Measurement error has been studied extensively in the survey literature (comprehensive reviews may be found in Groves 1989; Biemer and Lyberg 2003; and Groves et al. 2009). For many surveys, measurement error is one of the most damaging sources of error. It includes errors arising from respondents, interviewers, survey questions, and various interview factors. Respondents may (deliberately or unintentionally) provide incorrect information in response to questions. Interviewers can cause errors in a number of ways. By their speech, appearance, and mannerisms, they may undesirably influence responses, transcribe responses incorrectly, falsify data, or otherwise fail to comply with the survey procedures. The questionnaire can be a major source of error if it is poorly designed. Ambiguous questions, confusing instructions, and easily misunderstood terms are examples of questionnaire problems that can lead to measurement error.

However, measurement errors can also arise from the information systems that respondents may draw on to formulate their responses. For example, a farm operator or business owner may consult records that may be in error and thus cause an error in the reported data. It is also well known that the mode of administration can have a profound effect on measurement error (see, for example, Biemer and Lyberg 2003, Chapter 6; de Leeuw 2005). As an example, mode comparison studies (Biemer 1988; de Leeuw and van der Zouwen 1988) have found that data collected by telephone interviewing are, in some cases, less

accurate than the same information collected by face-to-face interviewing. Finally, the setting or environment within which the survey is conducted can also contribute to measurement error. For example, for collecting data on sensitive topics such as drug use, sexual behavior, or fertility, a private setting, even if using a self-response mode, is often more conducive to obtaining accurate responses than one in which other members of the household are present. In establishment surveys, topics such as land use, loss and profit, environmental waste treatment, and the allocation of corporate resources can also be sensitive. In these cases, assurances of confidentiality may reduce measurement errors due to intentional misreporting.

FRAME ERROR

Frame error arises in the process for constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. The sampling frame is defined as a list of target population members or another mechanism used for drawing the sample. Ideally, the frame would contain every member of the target population with no duplicates. Units that are not part of the target population would be removed from the frame. Likewise, information on the frame that is used in the sample selection process should be accurate and up to date. Unfortunately, sampling frames rarely satisfy these ideals, often resulting in various types of frame errors. In many situations, the most serious of these is frame omissions that lead to population *noncoverage* errors. An excellent discussion of frame error can be found in Lessler and Kalsbeek (1992).

NONRESPONSE ERROR

Nonresponse error is a fairly general source of error encompassing both unit and item nonresponse. Unit nonresponse error occurs when a sampled unit (e.g., household, farm, establishment) does not respond to any part of a questionnaire (e.g., a household that refuses to participate in a face-to-face survey, a mailed survey questionnaire that is never returned, an eligible sample member who cannot be contacted). Item nonresponse error occurs when the questionnaire is only partially completed because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. For example, income questions are typically subject to a high level of item nonresponse because of respondent refusals. Groves and Couper (1998) provides a comprehensive examination of the issues for nonresponse error in surveys.

DATA-PROCESSING ERROR

Data-processing error includes errors in editing, data entry, coding, assignment of survey weights, and tabulation of the survey data. As an example of editing error, suppose that a data editor is instructed to call back the respondent to verify the value of some budget-line item whenever the value of the item exceeds

a specified limit. In some cases, the editor may fail to apply this rule correctly, thus leaving potential errors in the data uncorrected.

The survey weights that statistically compensate for unequal selection probabilities, nonresponse errors, and frame coverage errors may be calculated erroneously, or there may be programming errors in the estimation software that computes the weights. Errors in the tabulation software may also affect the final data tables. For example, a spreadsheet used to compute the estimates may contain a cell-reference error that goes undetected. As a result, the weights are applied incorrectly and the survey estimates are in error. Chapter 7 in Biemer and Lyberg (2003) describes the various types of data-processing error, their effects on survey estimates, and how they can be controlled in surveys.

MINIMIZING TSE

Making the correct design decisions requires simultaneously considering many quality and cost factors and choosing the combination of design features and parameters that minimizes the TSE within all the specified constraints. To aid the design process, it is important to have a means of quantifying the total error in a survey process. That way, alternative survey designs that satisfy the specified constraints can be compared using their TSE as a criterion for determining the best design.

As an example, consider two survey designs—design A and design B—and suppose that both designs satisfy cost and other constraints for the survey. However, for the key characteristics to be measured in the study, the total error in the estimate for design A is 20 percent less than the TSE for design B. Obviously, the best design choice is design A, assuming other factors are equalized. Thus, the ability to summarize and quantify the total error in a survey process provides a method for choosing between competing designs.

A measure of TSE could also aid in the allocation of survey resources to minimize survey error. As an example, suppose we could establish that a major source of survey error for some design is due to nonresponse. This would suggest that efforts to further improve the quality of the survey data for this design should focus on reducing the effects of nonresponse on the data. Survey resources could then be reallocated in the design, if necessary, to better minimize the effects of nonresponse. This strategy will move the design closer to optimality if the overall effect is a reduction in the TSE. For example, shifting resources originally allocated to frame construction to nonresponse follow-up could reduce TSE even though frame error would be increased.

Mean Squared Error

Although a number of acceptable metrics for quantifying TSE have been proposed in the statistical literature, the most common metric for survey work is the *mean squared error* (MSE). Each estimate that will be computed from the

survey data has a corresponding MSE that summarizes the effects of all sources of error on the estimate. A small MSE indicates that the TSE is small and under control. A large MSE indicates that one or more sources of error are adversely affecting the accuracy of the estimate.

One of the primary uses of the MSE is as a measure of the accuracy of survey data. Unfortunately, it is seldom possible to compute the MSE directly in practical situations because this usually requires an estimate of the parameter that is essentially error free. Still, the concept is quite useful for understanding how the combined effects of survey errors reduce estimation accuracy. In addition, survey designers may benefit from the knowledge of these concepts through a better understanding of how their design decisions affect the overall quality of the survey data.

In statistical terms, MSE is the expected squared difference between an estimate, $\widehat{\theta}$, and the parameter it is intended to estimate, θ, which may be written as

$$MSE(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2 \tag{1}$$

or, after decomposing it into terms for the squared bias and the variance, as

$$MSE(\widehat{\theta}) = B^2(\widehat{\theta}) + Var(\widehat{\theta}) \tag{2}$$

As depicted in figure 1, for the purposes of this article, MSE reflects the cumulative effects of all sampling and nonsampling error sources on the survey estimate. This point will be emphasized by preceding MSE by the word "total" to distinguish this definition from less comprehensive forms of the MSE.

Each error source may contribute a variable error, a systematic error, or both. Variable errors are reflected in the variance of the estimate, while systematic errors are reflected in the bias squared component. The bias and variance components may be further decomposed into process-level and even subprocess-level components to further pinpoint specific error sources and, hopefully, their root causes. Such decompositions can be quite helpful for designing surveys and targeting and controlling the major error sources during survey implementation. For error evaluations, the major components of the total MSE are estimated and combined according to the decomposition formulas to form an estimate of the total MSE.

Next, consider a simple model for decomposing the total MSE of a particular characteristic in the survey labeled y. Survey errors that arise from all the various error sources in a survey have a cumulative effect on the observed value of y. The errors may cause the observed value of y to be higher or lower than its true value for an individual. Mathematically, this can be written as

$$y_i = \mu_i + \varepsilon_i, \tag{3}$$

where y_i and μ_i are the observed and true values, respectively, for unit i, and ε_i represents the cumulative effect of all error sources for the ith unit. The error

may be positive for some individuals and negative for others. If the net effect of these errors over the sample is close to 0, the estimate $\hat{\theta}$ will be close to the parameter θ, apart from sampling error.

For example, suppose θ is the population mean, which for a simple random sample is estimated by the sample mean denoted by $\bar{y}$. If $E(\varepsilon_i)= 0$, where expectation is taken over both the response distribution and the sampling distribution, then $E(\bar{y}) = \mu$, the true population mean, and $\bar{y}$ is said to be *unbiased* for μ. The ε_i satisfying these assumptions are called *variable errors* since, as we shall see, they add variation to the observations, but not bias.

In other situations, survey errors may be *systematic* (i.e., the sum of the errors across a typical sample is not zero because either positive or negative errors are dominant). As an example, the measurement errors for socially undesirable characteristics, such as excessive alcohol consumption, tend to be negative because heavy drinkers tend to underreport their amounts consumed. In this situation, $E(\varepsilon_i) < 0$ (i.e., the expected value of the errors over response and sampling distributions is negative and the observations are said to be *negatively biased,* which means that alcohol consumption will be underestimated). Similarly, positive errors may dominate for socially desirable characteristics, such as church attendance, voting behavior, charitable giving, and safe-sex practices. The positive systematic errors result in estimates that are *positively biased.*

Let $E(\varepsilon_i) = B$ denote the expected value of the error in (3) and note that

$$E(\bar{y}) = \mu + B. \tag{4}$$

B is called the *bias* in the estimator $\bar{y}$ for estimating μ. The model in (3) may be rewritten as

$$y_i = \mu_i + B + e_i, \tag{5}$$

where $e_i = \varepsilon_i - B, E(e_i) = 0, \mathrm{Var}(e_i) = \sigma_e^2$, and $\mathrm{Var}(\mu_i) = \sigma_\mu^2$. If we further assume that the errors between any two units are uncorrelated (i.e., $Cov(e_i, e_{i'}) = 0$ for any two units i and i'), the *MSE* of $\bar{y}$ can be written as

$$\begin{aligned} MSE(\bar{y}) &= B^2 + \frac{\sigma_\mu^2 + \sigma_e^2}{n} \\ &= B^2 + \frac{1}{R}\frac{\sigma_\mu^2}{n}, \end{aligned} \tag{6}$$

where

$$R = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_e^2} \tag{7}$$

is the *reliability ratio*. Note that R reflects all sources of random error, not just those arising from the measurement process.

For interview-assisted surveys, the assumption of uncorrelated errors may not hold, because of the effect of interviewers on the errors. Some interviewers, by their mannerisms, appearances, interactions with respondents, methods of probing or providing feedback, and other characteristics, may have a tendency to elicit responses that are more positive (or more negative) than other interviewers. As an example, there is ample evidence that when the races of the interviewer and the respondent differ, questions about racial issues can be predictably biased (see, for example, Schaeffer 1980). There is also evidence that experienced interviewers are more inclined than less experienced ones to change the wording of the questions in ways that affect responses. (See Groves 1989, or more recently, Biemer and Lyberg 2003, 149–187, for a review of the literature on interviewer effects.)

Interviewer errors share some properties of both variable and systematic errors in that they are systematic within an interviewer's work assignment, but are uncorrelated across work assignments. We refer to these errors as *intra-interviewer correlated errors.*

Suppose there are I interviewers available for the survey, and assume that each interviewer ($i = 1,\ldots,I$) imparts a bias, say b_i', to the observations in his or her assignment for some survey item. Assume that b_i' is the same for all respondents in the ith interviewer's work assignment. Let ε_{ij} denote the error in the observation for the jth unit in the ith interviewer's assignment. Under these assumptions, the conditional expectation of ε_{ij} (given interviewer i) is $E(\varepsilon_{ij}|j) = b_i'$. For the unconditional expectation, $E(\varepsilon_{ij}) = B$ as before. Let $b_i = b_i' - B$ denote the centered interviewer bias terms, and write

$$y_{ij} = \mu_{ij} + B + b_i + e_{ij}, \tag{8}$$

where μ_{ij} is the true value of the characteristic, $e_{ij} = \epsilon_{ij} - B - b_i$, $E(e_{ij}) = E(b_i) = 0$, $\text{Var}(e_{ij}) = \sigma_e^2$, and $\text{Var}(b_i) = \sigma_b^2$. We further assume that

$$\begin{aligned} Cov(y_{ji}, y_{j'i'}) &= \sigma_\mu^2 + \sigma_b^2 + \sigma_e^2 \text{ if } i = i', j = j' \\ &= \sigma_b^2 \qquad\qquad \text{if } i = i', j \neq j' \\ &= 0 \qquad\qquad\quad \text{if } i \neq i'. \end{aligned} \tag{9}$$

Again, assuming the n units are selected by simple random sampling (SRS) from a large population, the MSE of $\bar{y}$ under this model is

$$MSE(\bar{y}) = B^2 + \frac{\sigma_\mu^2 + \sigma_e^2}{n} + \frac{\sigma_b^2}{I}, \tag{10}$$

which differs from (6) by the addition of the interviewer variance term, σ_b^2/I.

Note that the interviewer effects, b_i, are considered to be random variables in this formulation rather than fixed effects. This is because evaluations of interviewer error are usually more concerned with the effects of the interviewing

process generally on the survey results rather than with the I specific interviewers employed to conduct the interviews. The interviewers are regarded as a random sample of I interviewers selected from a large population of interviewers. Note that the correlation between any two units in the same interviewer assignment is

$$\rho_{\text{int}} = \frac{\sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2}, \tag{11}$$

which is referred to as the *intra-interviewer correlation coefficient.* This parameter may also be interpreted as the proportion of the total variance of an observation due to interviewer variance.

Now, assume for simplicity that each interviewer is assigned exactly $m = n/I$ (an integer) cases, where n is the total sample size. Further, if we redefine the reliability ratio in (7) to include the interviewer variance component as follows,

$$R_{\text{int}} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_\varepsilon^2}, \tag{12}$$

then (10) can be rewritten as

$$MSE(\bar{y}) = B^2 + \frac{\sigma_\mu^2}{nR_{\text{int}}}(1 + (m-1)\rho_{\text{int}}). \tag{13}$$

Although somewhat oversimplified, this form of the MSE is instructive in that it contains terms for bias (B^2), sampling variance (σ_μ^2/n), reliability (R_{int}), and intra-interviewer correlation (ρ_{int}). The term $\frac{\sigma_\mu^2}{nR_{\text{int}}}$ is the variance of $\bar{y}$ when there is no interviewer variance (i.e., $\sigma_b^2 = 0$). With interviewer variance, the variance is increased by the factor $(1 + (m-1)\rho_{\text{int}})$, sometimes referred to as the *interviewer design effect* (deff_{int}). More complex expressions for the MSE that are derived under less restrictive assumptions can be found in Lessler and Kalsbeek (1992).

Even a seemingly small amount of interviewer-correlated error can have a profound impact on the TSE. As an example, consider a survey such as the U.S. Current Population Survey (CPS), which has an average interviewer workload size of approximately $m = 50$. Biemer and Lyberg (2003) and Groves (1989) note that values of ρ_{int} between 0.01 and 0.05 are not uncommon in face-to-face surveys, and values as high as 0.1 have been observed for some data items. Assuming a moderate value of 0.03, the value of deff_{int} is $[1+(50-1)\times 0.03] = 2.47$ (i.e., the variance is increased by almost 1½ times as a result of interviewer variance!). Similar expressions for correlated error variance can be derived for coders, keyers, editors, crew leaders, and other survey personnel (see, for example, Biemer and Lyberg 2003).

Total Survey Error Design Principles

We separately consider the design and implementation phases of a survey in the application of the TSE paradigm. All surveys are based upon a design that, to some degree, specifies the questionnaire content and format, the sampling plan, data-collection protocols, interviewer hiring, training, supervision approaches, plans for post-survey processing, weighting and analysis, schedule for completion, and costs. In some cases, the survey implementation goes according to plan, but in most cases, especially for new surveys, the initial design must be modified as the survey processes are executed to compensate for unforeseen data-collection issues, unanticipated costs, and scheduling problems. This section describes some useful design principles related to the TSE paradigm. Section 6 will then address several error-reduction strategies that have been used successfully during implementation.

Whether or not it is explicitly referenced, the TSE concept has been applied to survey design for decades. For example, research on optimal design preceding the 1960 U.S. Decennial Census clearly indicated the cost-effectiveness and error-minimization properties of an all-mail census process. As a result, a mail census protocol was adopted in preference to a face-to-face interviewer-assisted approach as a means of reducing TSE while minimizing data-collection costs (Eckler 1972, 105). Today, most large-scale surveys are designed to achieve objectives related to cost minimization, error reduction, and timeliness.

As previously noted, optimal survey design attempts to minimize the total MSE within specified cost (previously denoted by C_A) and timeliness constraints. In practice, this is quite a difficult task because the survey designer lacks the critical information required for design optimization. For example, knowledge of the contributions to TSE of each major error source is seldom available. Even if it were known, that information alone is insufficient because choosing among the many design alternatives and methodologies requires knowledge of how the various design choices affect the total MSE. As an example, the designer might ask "where should additional resources be directed to generate the largest reduction on the MSE: extensive interviewer training for nonresponse reduction, greater nonresponse follow-up intensity, or by offering larger incentives to sample members to encourage participation?" Or, "should a more expensive data collection mode be used, even if the sample size must be reduced significantly to stay within budget?"

Fortunately, detailed knowledge on costs, errors, and methodological effects of design alternatives are not needed for every survey design for two reasons: (a) design robustness; and (b) effect generalizability. *Design robustness* refers to the idea that the total MSE of an estimator may not change appreciably as the survey design features change. In other words, the point at which the MSE is minimized is said to be "flat" over a fairly substantial range of designs. For example, it is well known that the optimum allocation of the sample to the

various sampling stages in multistage sampling is fairly robust to suboptimal choices (see, for example, Cochran 1977).

Effect generalizability refers to the idea that design features found to be optimal for one survey are often generalizable to other similar surveys; for example, similar topics, target population, data-collection modes, and survey conditions. As an example, Dillman's *tailored design method* (Dillman, Smyth, and Christian 2009) makes use of this principle for optimizing mail surveys. Similar approaches are now being developed for Internet surveys (Couper 2008; Dillman, Smyth, and Christian 2009). Through meta-analyses involving hundreds of experiments on surveys spanning a wide range of topics, survey methodologists have identified what appear to be the "best" combinations of survey design and implementation techniques for maximizing response rates, minimizing measurement errors, and reducing survey costs for these survey modes. Dillman's tailored-design method prescribes the best combination of survey design choices to achieve an optimal design for mail and Internet surveys that can achieve good results across a wide range of survey topics, target populations, and data-collection organizations.

Standardized and generalized optimal design approaches have yet to be developed for interviewer-assisted data-collection modes or for surveying most types of special populations, regardless of the mode. Nevertheless, there exists a vast literature covering virtually all aspects of survey designs for many applications. As an example, there is literature on the relationship between length of interviewer training, training costs, and interviewer variance (see, for example, Fowler and Mangione 1985). Whether these relationships are transferable from one survey to another will depend upon the specifics of the application (e.g., survey topic, complexity, target population). There is also a considerable amount of literature relating nonresponse reduction methods, such as follow-up calls and incentives to response rates, and in some cases, nonresponse bias (see Singer and Kulka 2002 for a review of the literature). Perhaps the TSE paradigm that led to a theory of optimal design of mail and Internet surveys may one day be employed in the development of a theory and methodology for optimal face-to-face or telephone survey design.

Real-time Costs and TSE Reduction Strategies

Despite careful planning, and even under ideal circumstances, surveys are seldom executed exactly as they were designed, for several reasons. First, the survey sample itself is random, which introduces a considerable amount of unpredictability into the data-collection process. There are also numerous other sources of random "shocks" during the course of a survey, such as personnel changes, especially among field interviewers (FIs), the weather at the data-collection sites, staffing issues, catastrophic events, and other unforeseen complications. Costs may be considerably higher than expected in some areas of the design, and indicators of data quality, such as response rates, frame coverage

rates, missing data rates, and interviewer performance measures, may suggest that survey quality is faltering. It may be necessary to change the data-collection mode for some sample members or to introduce other interventions to deal with problems as they arise. A proactive, dynamic, flexible approach to survey implementation is needed to deal with these uncertainties.

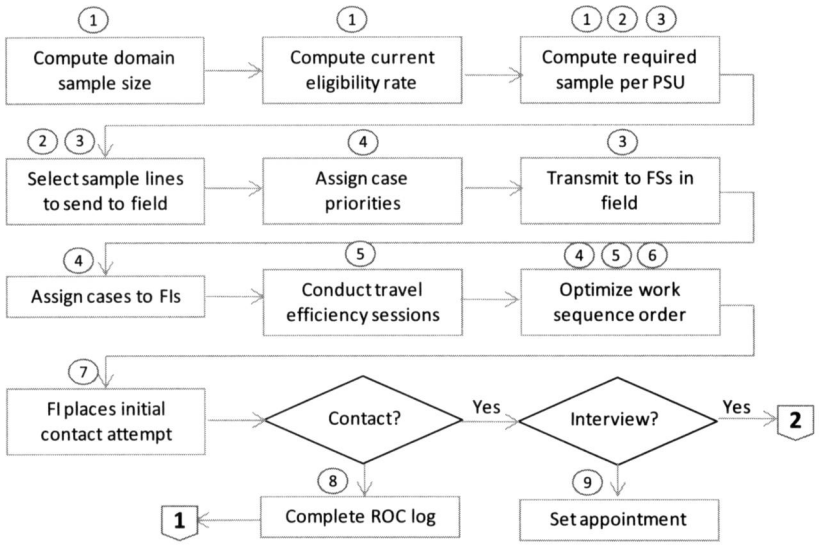
Thus, an essential ingredient of an optimal survey design is a plan for continuously monitoring key cost metrics and error-sensitive quality indicators to allow survey managers to control costs and reduce errors in real time. Real-time quality and production monitoring has always been an essential and integral part of survey implementation. However, with the advent of computer-assisted interviewing and computerized data processing, opportunities for collecting and monitoring process data (or *paradata*) have become proliferate. Within the last two decades, more structured and systematic strategies for quality monitoring have been devised that take advantage of the massive amounts of paradata generated by survey processes and the speed with which these data can be compiled, analyzed, reported, and visualized. Several important strategies for cost and error control during survey implementation are described in this section.

An approach that can be applied to virtually any survey operation is the *continuous quality improvement* (CQI) approach (Biemer and Caspar 1994; Morganstein and Marker 1997). A number of statistical organizations have adopted at least some aspects of CQI to control costs and errors in their surveys, including the U.S. Census Bureau (U.S. Department of Labor, Bureau of Labor Statistics, and U.S. Department of Commerce, Bureau of the Census 2002), Statistics Sweden (Lyberg 1985), Statistics Canada (Statistics Canada 2002), and Eurostat (Eurostat 2007). CQI uses a number of standard quality-management tools, such as the workflow diagram, cause-and-effect (or fishbone) diagram, Pareto histograms, statistical process control methods, and various production-efficiency metrics (see, for example, Montgomery 2009).

The CQI approach consists essentially of six steps, as follows:

1. Prepare a workflow diagram of the process and identify key process variables.
2. Identify characteristics of the process that are critical to quality (CTQ).
3. Develop real-time, reliable metrics for the cost and quality of each CTQ.
4. Verify that the process is stable (i.e., in statistical control) and capable (i.e., can produce the desired results).
5. Continuously monitor costs and quality metrics during the process.
6. Intervene as necessary to ensure that quality and costs are within acceptable limits.

The process workflow diagram (Step 1) is a graphical representation of the sequence of steps required to perform the process, from the initial inputs to the final output. In addition to the steps required, the flowchart can include a time line showing durations of activities, as well as annotations regarding inputs,



Critical to quality key:

1. Achieve target sample sizes for each domain
2. Distribute sample to PSUs to minimize design effects
3. Ensure that FI workloads are adequate
4. Achieve a high response propensity for high-priority cases
5. Minimize FI travel costs through work sequence optimization
6. Ensure that high-priority cases are worked fully
7. Ensure good cooperation at first contact
8. Ensure that record of calls (ROC) is completed accurately
9. Schedule a firm appointment after each contact

Figure 2. A Workflow Diagram for Sampling and the Initial Interview Attempt.

outputs, and activities that are deemed CTQ (Step 2). To illustrate, figure 2 shows a workflow diagram for selecting a monthly sample, sending it to the field, and conducting the initial attempt to interview the sample. A total of nine CTQs were identified for this part of the data-collection process. In Step 3, metrics were developed to monitor these CTQs during the various stages of the process. For example, to monitor whether domain target sample sizes were being achieved, the number of interviews per domain were compared to the required quota for the month. A model for predicting the final outcome of pending cases in the field was developed based upon similar cases whose interview status had been resolved.

Step 4 is important to establish that the quality of the process to be controlled (a) can be controlled; and (b) is capable of producing a product having acceptable cost and quality. A wildly erratic metric might suggest that it is unreliable and thus useless for quality monitoring and improvement. Unreliable metrics

can hide important deviations from acceptable quality or falsely indicate departures from acceptable quality. However, extreme variation in a reliable metric can be interpreted as an indication of process instability. Such processes cannot be easily improved until they are stabilized and capable of producing consistent results (referred to as *in a state of statistical process control*).

As an example, response rates for field interviewers (FIs) may vary widely, which can signal a problem with basic interviewer skills for some FIs. Retraining some FIs may solve this problem; otherwise, efforts to improve overall interviewing performance can be futile. Equally important is to establish that the process is capable of producing the desired results. For example, a data-collection process that consistently yields an average 50-percent response rate is in statistical process control. However, the process may be incapable of achieving a client-specified minimum 70-percent response rate without major changes to the process. CQI may be ineffective if the process is poorly designed and incapable, even under ideal circumstances, of producing the desired results.

The literature of statistical process control distinguishes between two types of process variation, referred to as *special cause* and *common cause*. *Special causes* are events or circumstances that are sporadic, unexpected, and traceable to somewhat unusual combinations of factors. As an example, weekly production drops in an area because the FI assigned to the area resigned in mid-data collection. Or productivity declines in the call center because a power outage stopped work for hours. Such problems are addressed by actions that are specific to the cause, leaving the design of the process essentially unchanged. By contrast, *common causes* are minor (chronic) disturbances that frequently and naturally occur during the normal course of the process. Such variations are inherent in the process and can be reduced only by redesigning the process. Specific actions to address common causes are not advisable because rather than reducing such variation, such actions (referred to as "tampering") may actually increase common cause variation. In some cases, it can even spawn more serious, special causes. As an example, FI costs and production rates normally fluctuate from period to period as a result of many uncontrollable variables, including the workload size, types of sampling units, FI behavior patterns, and the random locations of the units. Singling out FIs who happen to have low-response rates in any given period that are attributable to common cause variation can result in low staff morale and higher staff turnover.

The quality-control literature provides a number of tools for distinguishing between special and common cause variation. Chief among these are *control limits,* which define the boundaries of normal or routine variation of a metric. Control limits are set based upon previous values of the metric. For example, the lower control limit (LCL) and upper control limit (UCL) of a process can be computed for the metric (x) using the formulas $\text{LCL} = \bar{x} - 3\sigma$ and $\text{UCL} = \bar{x} + 3\sigma$, where $\bar{x}$ is the mean and σ is the standard deviation of x, both of which can be computed from the most recent 20–30 values of x. The range between the LCL and UCL of a process is 6σ, which, if the process is stable and

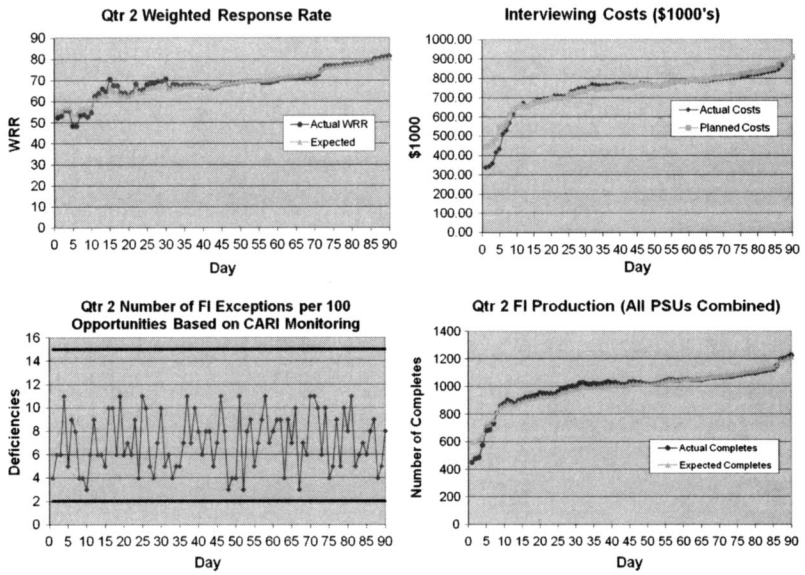


Figure 3. Illustration of a Dashboard Showing Weighted Response Rates, Interview Costs, FI Exceptions, and Production.

x can be assumed to be normally distributed, will bound x in about 99.7 percent of the cases. This means a value of x outside the control limits is likely due to a special cause. Otherwise, fluctuations within the control limits should be regarded as common cause (or random) variation. The third panel in figure 3 (which is described more fully below) is an illustration of control limits for an FI exception report. Note that the variation, while large, is still within control limits. Reducing this variation is better accomplished by redesigning the process rather than through specific FI-level behavioral changes, which will be largely ineffective.

In Step 5, decisions are made regarding the frequency for computing and reviewing metrics and the format of the data displays (e.g., tabular, graphical, descriptive). For some processes, it may be both informative and practicable to display related metrics together as a *dashboard*. Like the dashboard in an automobile, the CQI dashboard organizes and displays critical information on costs, timeliness, and quality across a wide spectrum of processes in a way that is easy to read and interpret. Unlike an automobile's dashboard, management should be able to interact with dashboards; for example, to modify the data displays to reveal different perspectives or views of the same data source in order to search for causalities. This is particularly useful for understanding why important metrics deviate from their benchmarks. Dashboards also provide the ability to "drill down" into the data in order to look for root causes and to investigate the effects of prior interventions and remedial actions.

As an example, figure 3 shows a typical high-level dashboard for monitoring response rates, FI production, FI performance quality, and costs for a quarterly survey. Each chart on the dashboard represents a "roll-up" of detailed, unit-level (e.g., case, FI, call attempt) data. As an example, the "FI exception report" in the figure is a process control chart plotting the number of issues detected during the routine monitoring of interviews using a system of digital recorded interviews on the FIs' laptops referred to as CARI (Computer Assisted Recorded Interviewing). An "FI exception" could be defined in a number of ways; for example, a major wording change, the use of inappropriate probes or feedback, or providing inaccurate information in response to respondent queries. Control limits on the chart suggest that the variation in the number of exceptions is likely due to common causes and no action is required. Data on the graphs can be hyperlinked to data at lower levels, for example, to identify the survey questions or the FIs that are the largest contributors to total exceptions. Special cause deviations can be investigated using this drill-down capability. Similar dashboards can be constructed for other survey processes, such as survey data processing and file preparation. It is important that the appropriate survey staff have online access to these dashboards to examine the metrics of greatest interest to them on an ongoing basis and to facilitate planning discussions.

The key to CQI is controlling TSE through the application of effective interventions at critical points during the process to address special causes (Step 6). In addition, process improvements can be implemented to reduce common cause variation and to improve the process average, $\bar{x}$. The error control and process improvement aspects of CQI tend to be the most challenging because of the knowledge and skill required to be effective. Process interventions must be timely and focused. Process improvements may require considerable time and repetition of the process (process *cycles*). In some cases, experimentation may be desirable. All the while, costs and timeliness must be held to strict constraints. *Responsive design* is an important innovation for better accomplishing these objectives.

Responsive design (Groves and Heeringa 2006) is a strategy developed for face-to-face data collection that includes many of the ideas, concepts, and approaches of CQI. It provides several additional and important new concepts and strategies that are intended to increase quality-monitoring sensitivity, data-collection efficiency, and intervention effectiveness. Like CQI, responsive design seeks to identify features of the survey design that are critical to data quality and costs and then to create valid indicators of the cost and error properties of those features. These indicators are closely monitored during data collection, and interventions are applied, as necessary, to reduce survey errors (primarily nonresponse bias) and costs.

What is unique is that responsive design organizes survey data collection around three (or more) phases: (1) an experimental phase, during which alternate design options (e.g., levels of incentives, choice of modes) are tested; (2) the main

data-collection phase, where the bulk of the data is collected using the design option selected after the first phase; and (3) the nonresponse follow-up phase, which is aimed at controlling costs and minimizing nonresponse bias, for example, by subsampling the nonrespondents from the second phase (i.e., *double sampling*), shifting to a more effective mode, and/or using larger incentives.

Responsive design recognizes that, in many cases, the survey designer is unable to choose the optimal data-collection approach from among several promising alternatives without extensive testing. In fact, it is common in survey work for data collection to be preceded by a pretest or pilot survey designed to identify the best data-collection strategy. Responsive design formalizes this practice and includes it as an integral part of the survey design. Another key concept of responsive design is the notion of a *phase capacity*. The main data-collection phase is said to have reached its phase capacity when efforts to reduce nonresponse and its biasing effects on selected survey estimates are no longer cost-effective. For example, after many attempts to follow up with nonrespondents, the key survey estimates remain unchanged and the data-collection phase is said to have reached its phase capacity. According to Groves and Heeringa (2006), a phase capacity condition signals the ideal point at which the main data-collection phase should be terminated and the third phase should begin.

The third phase intensifies the nonresponse follow-up operation from the second phase. However, to control costs, only a subsample (i.e., double sample) of the phase two nonrespondents are pursued in this phase. Nonrespondents that are not selected for the double sample are no longer pursued. A weight adjustment is applied to the nonrespondents who eventually respond in the third phase to represent the nonsampled nonrespondents. The subsample selection probabilities are typically a function of predicted response propensities, costs per follow-up attempt, the original case-selection weights, and projected sample design effects. Groves and Heeringa (2006) discuss a number of innovative metrics based upon paradata that can be used for CQI in all three phases, as well as approaches for determining when phase capacity has been reached.

Although responsive design focuses on nonresponse error, it can be combined with the TSE reduction strategies of CQI to provide a more comprehensive strategy for controlling costs and error. For example, as shown in Kreuter, Müller, and Trappmann (2010) and Kaminska, McCutcheon, and Billiet (2010), both in this issue, nonresponse reduction efforts can increase measurement errors. This might occur, for example, as a result of respondent satisficing (Krosnick and Alwin 1987) or interviewers who sacrifice data quality to avoid breakoffs (Peytchev, Peytcheva, and Groves 2010). Likewise, subsampling nonrespondents in the third phase may reduce the nonresponse bias, but can also substantially reduce the precision of the estimates as a consequence of increased weight variation (i.e., the unequal weighting effect; see, for example, Singh, Iannacchione, and Dever 2003). The usual method for controlling this variation is to trim the weights, but this can increase the estimation bias (see, for example, Potter 1990). Cumulatively, TSE could be substantially increased

even as the bias due to nonresponse bias is reduced. These risks to TSE are ameliorated by monitoring and controlling multiple sources simultaneously.

For this purpose, an even richer set of strategies for CQI can be found in the literature of Six Sigma (see, for example, Breyfogle 2003). Developed at Motorola in the early 1980s, Six Sigma embodies a set of principles and strategies for improving any process. Like CQI, Six Sigma emphasizes decision-making based on reliable data that are produced by stable processes, rather than intuition and guesswork. An important distinction between CQI and Six Sigma is the emphasis by the latter on providing verifiable evidence that quality improvements have been successful in improving quality and reducing costs, and that these gains are being held or further improved. Similar to the six steps outlined above for CQI, Six Sigma operates under the five-step process referred to as DMAIC: define the problem, measure key aspects of the process (i.e., CTQs) and collect relevant data, analyze the data to identify root causes, improve or optimize the current process using a set of Six Sigma tools designed for this purpose, and control and continue to monitor the process to hold the gains and effect further improvements.

We believe the most effective strategy for real-time survey cost and error reduction combines the phase-based approach of responsive design for controlling nonresponse error with the more general approaches of CQI and Six Sigma to simultaneously control all major sources of TSE. In particular, dashboards can be created based upon paradata to simultaneously monitor sampling error, nonresponse, measurement errors, and frame coverage errors during data collection, as suggested by figure 3. This would enable the survey manager, for example, to consider the effects of nonresponse reduction methods on these other error sources. Later, in the data-processing stage, additional metrics can be developed and continuously monitored to improve the data capture, editing, coding, and data file-preparation processes. This would allow the survey designer to be responsive to costs and errors throughout the survey process and across all major sources of TSE.

Total Survey Error Evaluation

A post-survey evaluation of at least some components of the total MSE is an essential part of the TSE paradigm. Standard errors for the estimates have been routinely reported for surveys for decades and are now considered essential documentation. Evaluations of nonsampling error components of the MSE are conducted with much less frequency. One exception is the analysis of nonresponse bias required by the U.S. Office of Management and Budget (OMB) for government-sponsored surveys that achieve response rates less than 80 percent (OMB 2006). While this focus on the nonresponse bias is welcome, there are still no requirements or guidelines for evaluating other components of the total MSE that are potentially more problematic for many uses of the data.

Nonsampling error evaluations address several dimensions of total survey quality. As noted in Section 3, they are essential for optimizing the allocation of resources in survey design to reduce the error contributed by specific processes. In experimentation, error evaluations are needed to compare the accuracy of data from alternative modes of data collection or estimation methods. Estimates of nonsampling errors (e.g., nonresponse bias analyses, measurement reliability studies) also provide valuable information to data users about data quality. Such evaluations can be important for understanding the uncertainty in estimates, for interpreting the results of data analysis, and for building confidence and credibility in the survey results.

This section provides a brief overview of methods for estimating the total MSE and its components. Because the purpose of this section is primarily pedagogical, simple random sampling will be assumed, although extensions to complex survey settings are available for all the methods (see, for example, Wolter 2007, Appendix D). Because of space limitations, only a few examples of evaluation studies are discussed for each error source. For a more comprehensive treatment of the topic, see Lessler and Kalsbeek (1992) and the additional references provided for specific MSE components.

TOTAL MEAN SQUARED ERROR ESTIMATION

For the rare situation where *gold standard* (i.e., practically error-free) measurements are available for every unit in the sample (including nonrespondents), the MSE (excluding the frame error component) can be estimated directly. Data that have been used in gold standard evaluations, including administrative records such as birth certificates, government tax records, population and government welfare registers, police records, or company records on number of employees, can sometimes be considered essentially error-free for evaluation purposes. A number of studies have attempted to obtain gold standard measurements from reconciled reinterview surveys (see Forsman and Schreiner 1991); in-depth, probing reinterviews (see Biemer 1988); or the collection of blood, urine, hair, or other biological specimens (Harrison 1997).

Gold standard measurements can be very difficult and costly to obtain in practice and may still be poor measurements. Research has also shown that administrative records data can be quite inaccurate and difficult to use (Jay, Belli, and Lepkowski 1994; Marquis 1978) as a result of differences in time reference periods and operational definitions, as well as errors in the records themselves. A number of articles show that reconciled reinterview data can be as erroneous as the original measurements they were intended to evaluate (see, for example, Biemer and Forsman 1992; Biemer et al. 2001; Sinclair and Gastwirth 1996). Even biological measures, such as hair analysis and urinalysis used in studies of drug use, contain substantial false-positive and false-negative errors for detecting some types of drug use (see, for example, Visher and

McFadden 1991). Still, useful approximations of the total MSE and valuable insights regarding nonsampling error can still be obtained through the use of these approaches.

Suppose both interview and gold standard measurements are available on all respondents and nonrespondents. Let $\bar{y}_R$ denote the mean of the survey responses, and let $\bar{\mu}$ denote the mean of the gold standard measurements for all sample units (including nonrespondents). Then the estimator of the bias in $\bar{y}_R$ is

$$\widehat{B} = \bar{y}_R - \bar{\mu}, \tag{14}$$

and further, an approximate estimator of the MSE of $\bar{y}_R$ is

$$\widehat{MSE}(\bar{y}_R) \doteq \widehat{B}^2 - v(\bar{\mu}) + 2\sqrt{v(\bar{y}_R)v(\bar{\mu})}, \tag{15}$$

where $v(\bar{y}_R)$ and $v(\bar{\mu})$ are variance estimators for $\bar{y}_R$ and $\bar{\mu}$, respectively (see Potter 1990). A similar formula holds for complex sampling. Note that because $\bar{y}_R$and $\bar{\mu}$ are based on the same sample, any frame bias will not be reflected by this estimator. An estimator of frame bias can be constructed as follows.

FRAME BIAS

Estimating frame undercoverage bias requires an estimate of the noncovered subpopulation mean denoted by $\bar{y}_{NC}$, as well as a measure of the relative size of the noncovered subpopulation. Let $\widehat{\gamma}_{NC}$ denote the estimate of the proportion of the target population missing from the frame (i.e., the noncoverage rate). Let $\bar{y}_C$ denote the sample mean, which, by definition, estimates the covered population mean. Then it can be shown that an estimator of the frame bias is

$$\widehat{B}_{NC} = \widehat{\gamma}_{NC}(\bar{y}_C - \bar{y}_{NC}) \tag{16}$$

(i.e., frame undercoverage bias is the product of the noncoverage rate and the difference between the mean of the covered and uncovered subpopulations).

It is apparent from that if the noncoverage rate is very small, the bias estimate will be small, no matter how large the difference is between the covered and noncovered subpopulations. As the noncoverage rate increases, the bias increases, but the rate of increase depends on the extent to which units on the frame differ from units that are missing from the frame. If the difference $\bar{y}_C - \bar{y}_{NC}$ is small, the bias will still be small.

Obtaining the estimate can be quite problematic and costly because it might entail accessing data sources that were not available during the frame construction process. As an example, for evaluating the coverage bias for a mail list frame, Iannacchione, Staab, and Redden (2003) used the *half-open interval* method. For this method, the addresses on the frame are first sorted in geographically proximal order, and a random sample of units is selected from the sorted frame. FIs are

instructed to search for missing frame units in the interval between a selected unit and the next unit on the frame; for example, the units between 1230 Elm Street (the selected unit) and 1240 Elm Street (the next frame unit), if any. New units discovered by this approach are then used to construct the estimator $\bar{y}_{NC}$.

NONRESPONSE BIAS

A similar bias formula applies for evaluating the magnitude of bias due to nonresponse. Suppose an estimate of the mean of the nonresponding units, denoted by $\bar{y}_{NR}$, is available. Let $\hat{\gamma}_{NR}$ denote the nonresponse rate. Let $\bar{y}_R$ denote the mean of the respondents to the survey. Then it can be shown that an estimator of the nonresponse bias is

$$\hat{B}_{NR} = \hat{\gamma}_{NR}(\bar{y}_R - \bar{y}_{NR}). \tag{17}$$

Although there has been much focus on nonresponse rates in the past, clearly shows that nonresponse bias is not just a function of the nonresponse rate, but also depends upon the difference between respondents and nonrespondents for the characteristics of interest. If the nonrespondents are not much different from the respondents for these characteristics, then the nonresponse bias might be quite small, even though the nonresponse rate is high.

To compute $\bar{y}_{NR}$, the characteristic y must be known for at least a sample of nonrespondents to the survey. This typically involves a nonresponse follow-up study where further efforts to interview nonrespondents are attempted using a preferred approach (e.g., more personal mode, higher incentive, more intensive contacting or tracing efforts). These more successful methods will produce data on a subsample of nonrespondents that can be used to compute $\bar{y}_{NR}$. Variables on the sampling frame that are highly correlated with y can also be used to evaluate$\bar{y}_{NR}$. For example, suppose y is "health insurance coverage," which is not on the frame. If income or some other variable correlated with y is available, it can be used as a proxy for y for the purposes of evaluating $\hat{B}_{NR}$. Groves and Couper (1998) provide an excellent overview of methods for estimating $\hat{B}_{NR}$.

MEASUREMENT ERROR

The estimation of measurement bias and variance also requires supplemental information that is not usually available from the main survey. The component most frequently estimated is the reliability ratio, which is typically estimated using a *test-retest* design. Suppose that the characteristic y is measured on two occasions, and further, that the true value, μ_i , does not change between occasions. Let y_{1i} denote the observed characteristic of the ith unit from the main survey, and let y_{2i} denote the second observation for the unit. Assume that the second measurement process independently replicates the error distribution of the main survey (i.e., assume that the errors, ε_{1i} and ε_{2i}, are independent and identically distributed). Under these assumptions, y_{1i} and y_{2i} are called *parallel* measurements. It can be shown that an estimator of R is given by

$$\widehat{R} = \frac{\sum_{i=1}^{n}(y_{1i} - y_{2i})^2}{s_1^2 + s_2^2}, \tag{18}$$

where $s_1^2 = \sum(y_{1i} - \bar{y}_1)^2/(n-1)$, $\bar{y}_1$ is the mean of y_{1i}, and s_2^2 is defined analogously for y_{2i}.

Alternative estimators of R have been used, particularly for the case of categorical variables. For example, Cohen's kappa (Cohen 1960)—or equivalently, the index of inconsistency (U.S. Department of Commerce, Bureau of the Census 1985)—is frequently used for estimating R for dichotomous variables. Biemer (2011) reviews a number of methods for estimating R for categorical data, particularly latent class analysis.

INTERVIEWER VARIANCE STUDIES

Estimating interviewer variance can be quite challenging from an operational perspective, particularly for face-to-face surveys. This is because the estimation process requires that households be randomly assigned to interviewers, a process called *interpenetration* (Mahalanobis 1946). Failure to interpenetrate interviewer assignments will result in biased estimators of interviewer variance. In face-to-face surveys, geographically proximate interviewer assignment areas may be combined so that the households in the combined area can be assigned at random to each interviewer working in that area. The interpenetration process is much simpler in centralized telephone surveys if the telephone numbers to be called during a particular shift are randomly assigned to all the interviewers working the shift.

One strategy for estimating ρ_{int} for face-to-face surveys is to interpenetrate proximate pairs of assignments to reduce the travel costs within the interpenetrated areas. Suppose K interviewer assignment pairs are formed and interpenetrated, and let $k = 1, 2, \ldots, K$ denote the kth interpenetrated pair. For simplicity, assume equal assignment sizes, m. (U.S. Department of Commerce, Bureau of the Census 1985 provides the formulas for unbalanced interpenetrated designs.) Let $\bar{y}_{kA}$ and $\bar{y}_{kB}$ denote means of the two assignments (i.e., for interviewers A and B) in the kth pair. Then an estimator of ρ_{int} is

$$\widehat{\rho}_{\text{int}} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{ms_{kb}^2 - s_{kw}^2}{ms_{kb}^2 + \frac{m-2}{2}s_{kw}^2}\right), \tag{19}$$

where s_{kA}^2 and s_{kB}^2 are the within-interviewer assignment variance estimates for $k = 1, \ldots, K$, $s_{kb}^2 = (\bar{y}_{kA} - \bar{y}_{kB})^2$, and $s_{kw}^2 = s_{kA}^2 + s_{kB}^2$ (U.S. Department of Commerce, Bureau of the Census 1985).

Using interpenetrating interviewer pairs in field studies is highly complex administratively because of increased interviewer travel costs, overlapping assignment areas, interviewer attrition, and other factors. However, for centralized telephone surveys, interpenetration is compatible with the way most

telephone centers randomly assign sample units to interviewers and, therefore, interpenetrating telephone interviewer assignments is much easier. The U.S. Department of Commerce, Bureau of the Census (1985) provides a detailed discussion of both the operational and technical issues associated with interpenetrated interviewer assignments.

MEASUREMENT BIAS

Traditionally, the estimation of measurement bias requires the existence of gold standard measurements for at least a random subsample of respondents. Let y_i and μ_i denote the sample interview and gold standard measurements, respectively, on the ith respondent. Then an estimator of the measurement bias is

$$B_{MEAS} = \frac{1}{n_r}\sum_{i=1}^{n_r}(y_i - \mu_i) = \tilde{y} - \tilde{\mu}, \text{ say,} \tag{20}$$

where n_r denotes the number of sample units for which both interview and gold standard data are available, and $\tilde{y}$ and $\tilde{\mu}$ are the means of these responses for the interview and reinterview, respectively. Biemer (2011) provides alternative estimators of the measurement bias in the case of categorical data focusing on estimates derived from latent class analysis.

DATA-PROCESSING ERROR

Many of the methods discussed previously for measurement variance and bias can also be applied to the estimation of data-processing error. For example, the estimation of the correlated error associated with operators (e.g., coders, editors, keyers) also requires interpenetrated work units or assignments, and the form of the estimator is the same as for interpenetrated assignment pairs. However, in an office environment, full interpenetration of operator assignments like that described above for estimating interviewer effects in centralized telephone surveys can be accomplished rather easily. To estimate the effect on total variance of systematic operator error, a random effects analysis of variance model could be used (see, for example, U.S. Department of Commerce, Bureau of the Census 1985). Likewise, estimation of operator bias (paralleling equation (20)) requires the use of either gold standard estimates or model-based approaches, such as those described by Biemer (2011).

Conclusions

Despite the important uses that estimates of TSE can fulfill, there are few examples of TSE studies in the literature. Two exceptions are Mulry and Spencer (1993) and Groves and Magilavy (1984). Quality profiles exist for only a few major surveys, including the CPS (Brooks and Bailar 1978), Survey

of Income and Program Participation (Kalton, Winglee, and Jabine 1998), U.S. Schools and Staffing Survey (Kalton et al. 2000), American Housing Survey (Chakrabarty and Torres 1996), and U.S. Residential Energy Consumption Survey (U.S. Energy Information Administration 1996). Quality reports that accompany survey results rarely report more than response rates, imputation rates, and perhaps other process metrics discussed in Section 6.

Although numerous studies of nonresponse bias have been reported, relatively less is known about other sources of nonsampling error. For example, interviewer variance is rarely estimated in centralized telephone surveys, even though the cost of doing so routinely is relatively small. Studies of frame bias or data-processing errors are seldom reported. Recently, Tourangeau, Groves, and Redline (2010) and Olsen (2006), as well as several articles in this volume, have investigated the relationship between propensity and measurement error with mixed results. The International Total Survey Error Workshops (ITSEW) were established in 2005 to encourage research on multiple error sources and their interactions.[1]

Smith (1990) and Platek and Särndal (2001) note a lack of progress over the last 50 years in integrating sampling and nonsampling error as measures of uncertainty. Indeed, routine reporting of nonsampling error components in surveys seems unlikely because evaluation studies are often operationally complex, expensive to implement, and difficult to analyze, and often require sophisticated statistical models. Resources for evaluating TSE are usually not available, except for very large, ongoing surveys. Even then, they may be sufficient to assess only one or two sources of error, such as nonresponse bias or test-retest reliability.

Despite the lack of studies of TSE, the development of the total MSE concept has changed our way of thinking about survey design. Total MSE provides a conceptual framework for optimizing surveys that can still be quite useful, even if information on the relative magnitudes of the errors is lacking. As an example, knowing that a specified data-collection mode is likely to produce biased data may be sufficient motivation to search for a less biasing mode. Likewise, knowing that some important error sources are not well represented in our measures of uncertainty should cause one to temper claims regarding statistical accuracy or precision of survey estimates. For reducing survey error, the idea of parsing the error into specific sources and then further subdividing these into smaller, more manageable sources is a much better strategy than less focused, impractical approaches aimed at generally reducing TSE. Finally, the TSE framework provides a useful taxonomy for the study of nonsampling error. In fact, the quality profile, which is based on this taxonomy, is useful for

1. Contents of the past four ITSEW workshops can be viewed at http://www.niss.org/event/niss-affiliates-workshop-total-survey-error-march-17-18-2005; http://www.niss.org/event/itsew-2008; http://www.niss.org/event/itsew-2009; and http://www.niss.org/events/itsew-2010.

accumulating all that is known about specific sources of error, but also for indicating where there are important gaps in our knowledge. For example, the quality profiles done to date suggest that specification errors, data-processing errors, and frame errors appear to be neglected in the study of nonsampling error (Doyle and Clark 2001; Kasprzyk and Kalton 2001).

There are a number of promising new approaches to the evaluation and analysis of TSE that have generated some excitement and the promise of greater activity in the field. These include the use of Bayesian models for measurement error (Zaslavsky 2005), multilevel modeling of interviewer effects under unequal probability multistage sampling (Hox 2005), meta-analysis of reliability and validity studies to inform survey design (Saris, van der Veld, and Gallhofer 2004), latent class modeling of survey error (Biemer 2010), and the use of structural equation models for studying reliability and validity (Saris and Gallhofer 2007).

Future directions for the field are numerous. Many surveys are now attempting to use resource-allocation models that explicitly consider other major components of the TSE besides sampling error. However, more research is needed in the realm of data-processing error, particularly editing error. Several recent papers have suggested that survey data are being *overedited* (i.e., editing to the detriment of both data-quality and optimal-resource allocation). There is a need for additional quality profiles, particularly for major government-data programs in the U.S. and elsewhere. The field would also benefit from understanding how multiple sources of survey error interact, so that as we attempt to reduce the error from one source, we do not inadvertently increase the error in one or more other error sources.

If the past is prologue to the future, survey research will face important challenges as costs continue to rise and quality continues to decline, especially as a result of greater nonresponse. Recent advances in computer-assisted interviewing, uses of paradata, and new, more effective CQI strategies are essential devices for offsetting these threats to total survey quality. Future research is needed in three critical areas: (1) innovative uses of paradata for monitoring costs and quality during survey implementation; (2) research on highly effective intervention strategies for real-time costs and error reduction; and (3) cost-effective methods for evaluating survey error, particularly error interaction effects such as the effects of nonresponse reduction strategies on measurement error.

References

Biemer, Paul. 1988. "Measuring Data Quality." In *Telephone Survey Methodology*, eds. Robert Groves, Paul Biemer, Lars Lyberg, James Massey, William Nicholls, and Joseph Waksberg. New York: John Wiley & Sons, 273–82.

———. 2004. "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions." *Survey Methodology* 30(2):127–40.

———. 2011. *Latent Class Analysis of Survey Error*. Hoboken, NJ: John Wiley & Sons.

Biemer, Paul, and Rachel Caspar. 1994. "Continuous Quality Improvement for Survey Operations: Some General Principles and Applications." *Journal of Official Statistics* 10:307–26.

Biemer, Paul, and Gosta Forsman. 1992. "On the Quality of Reinterview Data with Applications to the Current Population Survey." *Journal of the American Statistical Association* 87(420): 915–23.

Biemer, Paul, and Lars Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons.

Biemer, Paul, Henry Woltman, David Raglin, and Joan Hill. 2001. "Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis." *Journal of Official Statistics* 17(1):129–49.

Breyfogle, Forrest. 2003. *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. Hoboken, NJ: John Wiley & Sons.

Brooks, Camilla A., and Barbara A. Bailar. 1978. *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Working Paper 3. Washington, DC: U.S. Office for Management and Budget.

Chakrabarty, Rameswar P., and Georgina Torres. 1996. *American Housing Survey: A Quality Profile*. Washington, DC: U.S. Department of Housing and Urban Development and U.S. Department of Commerce.

Cochran, William G. 1977. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurements* 20:37–46.

Couper, Mick P. 2008. *Designing Effective Web Surveys*. New York: Cambridge University Press.

de Leeuw, Edith D. 2005. "To Mix or Not to Mix Data-collection Modes in Surveys." *Journal of Official Statistics* 21(2):233–55.

de Leeuw, Edith D., and Johannes van der Zouwen. 1988. "Data Quality in Telephone Surveys and Face-to-face Surveys: A Comparative Meta-analysis." In *Telephone Survey Methodology*, eds. Robert Groves, Paul P. Biemer, Lars Lyberg, James Massey, William Nicholls, and Joseph Waksberg. New York: John Wiley & Sons, 273–82.

Dillman, Don, Jolene Smyth, and Leah Christian. 2009. *Internet, Mail, and Mixed-mode Surveys: The Tailored-design Method*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Doyle, Pat, and Cynthia Clark 2001. *Quality Profiles and Data Users*. Paper presented at the International Conference on Quality in Official Statistics. Stockholm: Sweden.

Eckler, A. Ross. 1972. *The Bureau of the Census*. New York: Praeger.

Eurostat. 2007. *Handbook on Data Quality Assessment Methods and Tools*, eds. Manfred Ehling and Thomas Körner. http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20-DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf (accessed 12/11/2010).

Fellegi, Ivan P. 1996. "Characteristics of an Effective Statistical System." *International Statistical Review* 64(2):165–97.

Forsman, Gosta, and Irwin Schreiner 1991. "The Design and Analysis of Reinterview: An Overview." In *Measurement Errors in Surveys*, eds. Paul Biemer, Robert Groves, Lars Lyberg, Nancy Mathiowetz, SudmanSeymour. New York: John Wiley & Sons, 279–302.

Fowler, Floyd J., and Thomas W. Mangione. 1985. *The Value of Interviewer Training and Supervision*. Final Report to the National Center for Health Services Research, Grant No. 3-R18-HS04189.

Groves, Robert. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Groves, Robert, and Mick Couper 1998. *Household Survey Nonresponse*. New York: John Wiley & Sons.

Groves, Robert, Floyd J Fowler, Mick Couper, James Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Groves, Robert, and Steven Heeringa 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society Series A* 169(3):439–57.

Groves, Robert, and Lou Magilavy. 1984. "An Experimental Measurement of Total Survey Error." *Proceedings of the Survey Research Methods Section,* 698–703. American Statistical Association.

Harrison, Lana. 1997. "The Validity of Self-reported Drug Use in Survey Research: An Overview and Critique of Research Methods." In *NIDA Research Monograph 97-4147*, eds. Lana Harrison, HughesArthur, 167:17–36.

Hox, Joop. 2005. "Multilevel Models in Survey Error Estimation." Presented at the Workshop on Total Survey Error. Washington, DC. http://www.niss.org/event/niss-affiliates-workshop-total-survey-error-march-17-18-2005 (accessed 12/11/2010).

Iannacchione, Vincent, Jennifer Staab, and David Redden. 2003. "Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey." *Public Opinion Quarterly* 67(2): 202–10.

Jay, Gina M., Robert F. Belli, and James M. Lepkowski. 1994. "Quality of Last Doctor Visit Reports: A Comparison of Medical Records and Survey Data." In *Proceedings of the ASA Section on Survey Research Methods,* 362–67.

Juran, Joseph, and Frank Gryna. 1980. *Quality Planning and Analysis*. 2nd ed. New York: McGraw-Hill.

Kalton, Graham, Marianne Winglee, Sheila Krawchuk, and Daniel Levine. 2000. *Quality Profile for SASS: Rounds 1–3: 1987–1995. NCES 2000-308.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kalton, Graham, Mariane Winglee, and Thomas Jabine 1998. *SIPP Quality Profile*. 3rd ed. Washington, DC: U.S. Bureau of the Census.

Kaminski, Olena, Alan McCutcheon, and Jaak Billiet. 2010. "Satisficing Among Reluctant Respondents in a Cross-national Context." *Public Opinion Quarterly* 74:880–906.

Kasprzyk, Daniel, and Graham Kalton 2001. "Quality Profiles in U.S. Statistical Agencies." Paper presented at the International Conference on Quality in Official Statistics. Stockholm: Sweden.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research." *Public Opinion Quarterly* 74:985–1003.

Krosnick, Jon A., and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201–19.

Lessler, Judith, and William Kalsbeek. 1992. *Nonsampling Errors in Surveys*. New York: John Wiley & Sons.

Lyberg, Lars. 1985. "Quality Control Procedures at Statistics Sweden." *Communications in Statistics—Theory and Methods* 14(11):2705–51.

Lyberg, Lars, Sven Felme, and Lars Olsson. 1977. *Kvalitetsskydd av data (Data protection)*. Stockholm, Sweden: Liber (in Swedish).

Mahalanobis, Prasanta C. 1946. "Recent Experiments in Statistical Sampling in the Indian Statistical Institute." *Journal of the Royal Statistical Society* 109:325–78.

Marquis, Kent H. 1978. "Inferring Health Interview Response Bias from Imperfect Record Checks." In *Proceedings of the ASA Section on Survey Research Methods,* 265–70.

Montgomery, Douglas C. 2009. *Introduction to Statistical Quality Control*. 6th ed. Hoboken, NJ: John Wiley & Sons.

Morganstein, David R., and David A. Marker. 1997. "Continuous Quality Improvement in Statistical Agencies." In *Survey Measurement and Process Quality*, eds. Lars E. Lyberg, Paul Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: John Wiley & Sons, 475–500.

Mulry, Mary, and Bruce Spencer. 1993. "The Accuracy of the 1990 Census and Undercount Adjustments." *Journal of the American Statistical Association* 88:1080–91.

Office of Management and Budget. 2006. *Questions and Answers When Designing Surveys for Information Collections*. Washington, DC: Office of Information and Regulatory Affairs. OMB. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/statpolicy/standards_stat_surveys.pdf (accessed 12/11/2010).

Olsen, Kristen. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70:737–58.

Peytchev, Andy, Emilia Peytcheva, and Robert M. Groves. 2010. "Measurement Error, Unit Nonresponse, and Self-reports of Abortion Experiences." *Public Opinion Quarterly* 74(2):319–27.

Platek, Richard, and Carl-Erik Särndal. 2001. "Can a Statistician Deliver?" *Journal of Official Statistics* 17(1):1–20.

Potter, Frank J. 1990. "A Study of Procedures to Identify and Trim Extreme Survey Weights." In *Proceedings of the American Statistical Association, Survey Research Methods Section*, Anaheim, CA.

Saris, Willem E., and Irmtraud Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research.* New York: John Wiley & Sons.

Saris, Willem E., William van der Veld, and Irmtraud Gallhofer. 2004. "Development and Improvement of Questionnaires Using Predictions of Reliability and Validity." In *Methods for Testing and Evaluating Survey Questionnaires*, eds. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: John Wiley & Sons, 275–98.

Schaeffer, Nora Cate. 1980. "Evaluating Race-of-interviewer Effects in a National Survey." *Sociological Methods & Research* 8(4):400–419.

Sinclair, Michael, and Joseph Gastwirth. 1996. "On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data." *Journal of the American Statistical Association* 91:961–69.

Singer, Eleanor, and Richard A. Kulka. 2002. "Paying Respondents for Survey Participation" In *Improving Measurement of Low-income Populations.* Washington, DC: National Research Council, National Academies Press.

Singh, Avinash C., Vincent G. Iannacchione, and Jill A. Dever. 2003. "Efficient Estimation for Surveys with Nonresponse Follow-up Using Dual-frame Calibration." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 3919–30.

Smith, T. M. Fred. 1990. "Comment on Rao and Bellhouse: Foundations of Survey-based Estimation and Analysis." *Survey Methodology* 16:26–29.

Canada, Statistics. 2002. *Statistics Canada's Quality Assurance Framework—2002.* Ottawa, Ontario: Catalogue No. 12-586-XIE.

Tourangeau, Roger, Robert Groves, and Cleo Redline. 2010. "Sensitive Topic and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74(2):413–32.

U.S. Department of Commerce, Bureau of the Census. 1985. *Evaluating Censuses of Population and Housing, STD-ISP-TR-5.* Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor, Bureau of Labor Statistics, and U.S. Department of Commerce, Bureau of the Census. 2002. *Current Population Survey: Design and Methodology,* Technical Paper 63RV. Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics, and U.S. Department of Commerce, Bureau of the Census. http://www.census.gov/prod/2002pubs/tp63rv.pdf (accessed 10/24/10).

U.S. Energy Information Administration. 1996. *Residential Energy Consumption Survey Quality Profile.* Washington, DC: U.S. Department of Energy.

Visher, Christy, and Karen McFadden. 1991. "A Comparison of Urinalysis Technologies for Drug Testing in Criminal Justice." In *National Institute of Justice Research in Action.* Washington, DC: U.S. Department of Justice, National Institute of Justice.

Wolter, Kirk. 2007. *Introduction to Variance Estimation.* 2nd ed. New York: Springer-Verlag.

Zaslavsky, Alexander. 2005. "Bayesian Modeling of Nonsampling Error." Paper presented at the Workshop on Total Survey Error. Washington, DC. http://www.niss.org/event/niss-affiliates-workshop-total-survey-error-march-17-18-2005 (accessed 12/11/2010).