Public Opinion Quarterly, Vol. 74, No. 5, 2010, pp. 849–879

TOTAL SURVEY ERROR
PAST, PRESENT, AND FUTURE

ROBERT M. GROVES
LARS LYBERG*

Abstract ''Total survey error'' is a conceptual framework describing statistical error properties of sample survey statistics. Early in the history of sample surveys, it arose as a tool to focus on implications of various gaps between the conditions under which probability samples yielded unbiased estimates of finite population parameters and practical situations in implementing survey design. While the framework permits design-based estimates of various error components, many of the design burdens to produce those estimates are large, and in practice most surveys do not implement them. Further, the framework does not incorporate other, nonstatistical, dimensions of quality that are commonly utilized in evaluating statistical information. The importation of new modeling tools brings new promise to measuring total survey error components, but also new challenges. A lasting value of the total survey error framework is at the design stage of a survey, to attempt a balance of costs and various errors. Indeed, this framework is the central organizing structure of the field of survey methodology.

Introduction

Total survey error is a concept that purports to describe statistical properties of survey estimates, incorporating a variety of error sources. For many within the field of survey methodology, it is the dominant paradigm. Survey designers can

ROBERT M. GROVES is Director, U.S. Census Bureau, Washington, DC, USA. LARS LYBERG is Professor Emeritus, Stockholm University, Stockholm, Sweden. The authors thank Anne Morillon of the U.S. Government Accountability Office for providing a glossary of terms used by official statistics and statistical indicator systems, upon which the conceptual map in figure 1 is based. Robert Groves acknowledges partial support from the National Science Foundation [SES 020733], while affiliated with the Survey Research Institute, University of Michigan, Ann Arbor, MI, USA. The views expressed in this article are the authors' and do not necessarily reflect the views of the U.S. Census Bureau, the U.S. Department of Commerce, or Stockholm University. *Address correspondence to Lars Lyberg, Department of Statistics, Stockholm University, Stockholm, Sweden; e-mail: lars.lyberg@stat.su.se.

doi: 10.1093/poq/nfq065

use total survey error as a planning criterion. Among a set of alternative designs, the design that gives the smallest total survey error (for a given fixed cost) should be chosen.

The mathematical statistical theory underlying *sampling* variance properties of descriptive statistics from probability samples of finite populations was elucidated in the 1930s (Neyman 1934). However, most of the inferential properties of the sample statistics treated in that theory required strong assumptions about nonsampling errors. The theory laid out in Neyman's landmark paper and other papers written by him in the 1930s on various sampling schemes applies only when nonsampling errors are small. Early on in practical survey work, many of those assumptions were called into question.

Unfortunately, the term "total survey error" is not well defined. Different researchers include different components of error within it, and a number of typologies exist in the literature. "Total survey error" contains some constituent elements that can be estimated with relatively minor alterations of traditional sample survey design. Over the years, it has generally been acknowledged that sampling variance is measurable in most probability sample surveys, but that other components of the notion cannot be directly measured without significant alteration of the typical survey designs. Inherent in the term total survey error is attention to the entire set of survey design components that identify the population, describe the sample, access responding units among the sample, operationalize constructs that are the target of the measurement, obtain responses to the measurements, and summarize the data for estimating some stated population parameter.

The total survey error of an estimate can be viewed as an indicator of data quality usually measured by the accuracy or the mean squared error (MSE) of the estimate. Early researchers such as Kendall, Palmer, Deming, Stephan, Hansen, Hurwitz, Pritzker, Tepping, and Mahalanobis viewed a small error as an indication of survey usefulness, and Kish (1965) and Zarkovich (1966) were the first to equate a small error with the broader concept of survey data quality. Later, other researchers and statistical organizations developed frameworks that include not only accuracy but also nonstatistical indicators of data quality such as relevance, timeliness, accessibility, coherence, and comparability (e.g., Felme et al. 1976; Eurostat 2000; Brackstone 1999; and OECD 2003).

Although it is easy to see that such indicators can add value to a statistical product, these extended frameworks could have benefited from more user consultation. As a matter of fact, we know very little about how users perceive and use information about quality dimensions. One exception is a study conducted by Hert and Dern (2006). In a design situation, the nonstatistical dimensions can be viewed as constraints when using the total survey error planning criterion. We will not treat these nonstatistical dimensions in the remainder of this article except when they have a distinct bearing on the error typologies and error models discussed.

This article reviews the conceptual history of "total survey error," offers comments on its evolution, identifies its contributions to the development of the field of survey methodology, criticizes weaknesses in the conceptual framework, and outlines a research agenda for future work to enrich the value of the framework.

Evolution of the Concept of "Total Survey Error"

We begin with a bit of history on the development of the concept. Deming (1944), in an early article (published in a sociology journal!), first outlined multiple error sources in sample surveys as "classification of factors affecting the ultimate usefulness of a survey" (p. 359). From that start, the concept has been elaborated and enriched by scores of researchers. The history shows two evolutionary paths: (a) increasing categorization of errors on dimensions of variance and bias on one hand, and errors of observation and nonobservation on the other; and (b) a very slow acknowledgment of other fields' quality orientations.

EARLY EVOLUTION OF "TOTAL SURVEY ERROR"

Figure 1 contains the actual error typology from Deming's 1944 *American Sociological Review* article. The list of "13 factors that affect the usefulness

1. Variability in response;
2. Differences between different kinds and degrees of canvass;
   (a) Mail, telephone, telegraph, direct interview;
   (b) Intensive vs. extensive interviews;
   (c) Long vs. short schedules;
   (d) Check block plan vs. response;
   (e) Correspondence panel and key reporters;
3. Bias and variation arising from the interviewer;
4. Bias of the auspices;
5. Imperfections in the design of the questionnaire and tabulation plans;
   (a) Lack of clarity in definitions; ambiguity; varying meanings of same word to different groups of people; eliciting an answer liable to misinterpretation;
   (b) Omitting questions that would be illuminating to the interpretation of other questions;
   (c) Emotionally toned words; leading questions; limiting response to a pattern;
   (d) Failing to perceive what tabulations would be most significant;
   (e) Encouraging nonresponse through formidable appearance;
6. Changes that take place in the universe before tabulations are available;
7. Bias arising from nonresponse (including omissions);
8. Bias arising from late reports;
9. Bias arising from an unrepresentative selection of date for the survey, or of the period covered;
10. Bias arising from an unrepresentative selection of respondents;
11. Sampling errors and biases;
12. Processing errors (coding, editing, calculating, tabulating, tallying, posting and consolidating);
13. Errors in interpretation;
   (a) Bias arising from bad curve fitting; wrong weighting; incorrect adjusting;
   (b) Misunderstanding the questionnaire; failure to take account of the respondents' difficulties (often through inadequate presentation of data); misunderstanding the method of collection and the nature of the data;
   (c) Personal bias in interpretation.

Figure 1. Deming's Listing of 13 Factors That Affect the Usefulness of a Survey (1944).

of a survey" is quite different from some of the later treatments of the notion of "total survey error." It *does* include nonresponse, sampling, interviewer effects, mode effects, various other types of measurement errors, and processing errors. However, it omits a separate category for coverage errors. It adds a set of nonstatistical notions, for example, biases arising from sponsorship or "auspices." Deming includes what we would now call "timeliness," in "changes that take place in the universe before tabulations are available" and similar ideas. Finally, Deming addressed errors that arise at the estimation stage—"bias arising from bad curve fitting; wrong weighting; incorrect adjusting." Of particular interest here is the absence of the nomenclature of total survey error.

It is clear that some of these notions can be specified in statistical terms only with unusual effort and even then would yield a rather immeasurable specification of error sources for practical survey design. The notion of "usefulness of a survey" implies the recognition of a user. Indeed, it appears that Deming's purpose in the article is to caution survey researchers and users of surveys to consider more than just sampling errors in interpreting the results of surveys.

Finally, Deming seems to address an issue that arises in the minds of most students when first encountering the total survey error framework—"If surveys are subject to all these error sources, how could a survey ever yield useful results?" Deming notes in one of the last sections of the article that

> It is not to be inferred from the foregoing material that there are grounds for discouragement or that the situation is entirely hopeless with regard to the attainment of useful accuracy. My point is that the accuracy supposedly required of a proposed survey is frequently exaggerated—is in fact often unattainable—yet the survey when completed turns out to be useful in the sense of helping to provide a rational basis for action. (p. 369)

We must view Deming's statement about accuracy requirements being exaggerated in light of what was known at the time and his overarching goal of promoting surveys based on sampling as a means to collect accurate information. One of the key values of Deming's early contribution is his focus on bias components of error versus variance components of error. Thus, it is at this very point in the history that it is clear that the concept would not routinely reach a status of fully measured error sources, as bias terms require some "gold standard" or "true" value to achieve measurability. For example, for decades demographic analysis was used as a gold standard for U.S. population estimates, from which undercoverage bias of U.S. censuses was estimated. (In most practical survey work, if gold standard measurements are already available on a "representative" sample, a survey is not justified.)

The writings of Deming and other statistician founders of surveys had their greatest influence on succeeding generations through their textbooks. At that point in the history of surveys, most textbooks treating surveys were sampling texts. Deming's 1950 text, *Some Theory of Sampling*, does begin with a re-

presentation of his 1944 article of sources of various errors in surveys, but then focuses entirely on sampling error properties. This is not really surprising, though, since sampling was not universally accepted and still had to be vigorously promoted at the time. The founders were in a first-things-first situation. But it is fair to say that almost all the classic text-length treatments of sampling theory and practice included some acknowledgment of "nonsampling error." For example, the 638-page *Sample Survey Methods and Theory,* volume 1 (Hansen, Hurwitz, and Madow 1953), included nine pages, labeled "Response and Other Nonsampling Errors in Surveys," that identified the major error sources as "errors due to faulty planning or definitions," "response errors," "errors in coverage," "errors in classification," "compiling errors," and "publication errors." "Classification errors" are noted to be response errors in assigning sample cases to key domains, while "compiling errors" are similar to the processing errors of Deming (1944). "Publication errors" include failure to tell users of the limitations of data or other errors of misinterpretation.

In his 413-page text, *Sampling Techniques*, Cochran (1953) included a final chapter of about 40 pages, labeled "Sources of Error in Surveys," that reviewed nonresponse errors, various measurement errors, and "errors introduced in editing, coding, and tabulating the results." His treatment of nonresponse is in large part not original, depending heavily on the work of Deming and of Hansen and Hurwitz on double sampling. The presentation on measurement errors distinguishes correlated and uncorrelated response variances. Later, Cochran (1968) regretted that sampling texts, including his own, contained so little material on other error sources and their treatment.

Kish's *Survey Sampling* (1965) includes 65 of its 643 pages in a chapter labeled "Biases and Nonsampling Errors," a notable increase in emphasis, probably because Kish worked in an environment filled with social scientists greatly interested in measurement and nonresponse properties of surveys. Figure 2 presents the schematic of Kish (1965) that begins to define relationships among the error sources, unfortunately focusing entirely on biases.

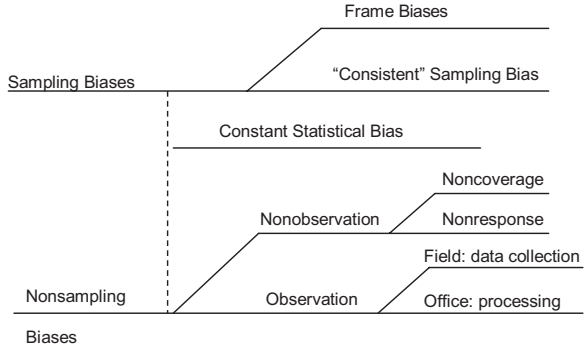

Figure 2. Schematic Presentation in Kish (1965) of Biases in Surveys.

His notion of "frame biases" at the top of the figure is distinguished from noncoverage biases in that frame conditions sometimes cause multiplicities of selection that require selection weighting. By "constant statistical bias," he meant properties of estimators, like the bias of the ratio mean to estimate a population mean or the biases of the median to estimate the mean of a skewed distribution. His is the first treatment to separate errors of observation from nonobservation, but he fails to note that sampling errors are inherently errors of nonobservation.

One of the purposes of Kish's and subsequent typologies of specific error sources is to list them and not forget important ones. Any listing is bound to be incomplete, though, since new error structures emerge due to new technology, methodological innovations, or strategic design decisions such as using mixed-mode data collection. All error sources are not known, some defy expression, and some can become important because of the specific study aim, such as translation error in cross-national studies. Having said that, it is still somewhat surprising that the typologies and the labeling of errors found in the literature vary so much between authors, and that they all have their own error-source-inclusion strategies. That might be an indication of a highly specialized survey research community.

The term "total survey design" was introduced by Dalenius in 1974 as a framework for an extensive research project, "Errors in Surveys," that he managed from 1967 until 1974. The term was based on Hansen, Hurwitz, and Pritzker's (1967) idea that there are three perspectives of a survey: requirements, specifications, and operations. Hansen, Hurwitz, and Pritzker claimed that corresponding to the requirements is a design, and if this design is properly carried out, the survey would yield a set of statistics, the ideal goals *Z.* The specifications are the actual design choices, and they identify the defined goals *X,* while the actual operations carried out yield the survey results *y.* Dalenius defined a total survey design process in which the design that comes closest to the defined goal is chosen. "Total" meant that one should be able to control all important error sources, and consequently "total" included the design of the survey operations as well as control and evaluation systems.

In 1979, Anderson, Kasper, and Frankel published *Total Survey Error*, a book whose title everyone in the field now covets, which presented a set of empirical studies about individual nonresponse, measurement, and processing error features of health survey data. The introductory chapter of the book displays an enhanced decomposition of total survey error, emphasizing principal divisions by variance and bias, then by sampling and nonsampling, then by observation and nonobservation. They omit the notion of consistent statistical bias arising from the inherent properties of the estimate (e.g., bias of the ratio mean).

Groves (1989) attempted to link error notions from the total survey error formulation with those of psychometrics and econometrics. He presented a nested structure within "mean squared error," first splitting on "variance" and "bias," then on "errors of nonobservation" and "observational errors." Some concepts in his structure are those that appear absent in psychometrics;

some other concepts are those covered by psychometrics but absent in survey statistics. The key messages of the discussion in the book were that (a) the psychometric notions of reliability were conceptually equivalent to those of simple response variance in the survey literature (e.g., O'Muircheartaigh 1991); but (b) the correspondence between notions of validity and survey error concepts was complex. First, definitions of validity were numerous (complicated by expansion of adjectives placed in front of "validity," like "statistical conclusion," "predictive," "concurrent," and "theoretical"). Second, it was clear that "validity" in psychometrics differed from "bias" in survey statistics, despite the colloquial equivalence of the terms.

The other contribution of the total survey error framework is its attention to the distinction between errors of nonobservation and errors of observation. Errors of nonobservation usually include coverage error (discrepancies on the survey statistics between the target population and the frame population), sampling error, and both unit and item nonresponse. Errors of observation involve differences between reported/recorded values of a survey variable and some "true" or underlying value. These errors of observation have produced a research literature that separately examines the influences of interviewers, mode of data collection, questionnaire, respondents, and post-survey data-processing alteration of reported data, as well as all the combined effects of those sources. These distinctions were present in Groves's formulation (1989), even though he omits processing error from his discussion.

In their 1992 volume, *Nonsampling Error in Surveys,* Lessler and Kalsbeek structure their presentation about frame errors, sampling errors, nonresponse errors, and measurement errors. It is important to note that Lessler and Kalsbeek evoke more explicitly the notion of "total survey design" than "total survey error" and clearly were influenced by the same notion as Dalenius that an appropriate goal was designing surveys with multiple error sources in mind.

In their 2003 volume, *Introduction to Survey Quality*, Biemer and Lyberg appear to force attention to the major division of sampling and nonsampling error, then list five major sources of nonsampling error: specification error, frame error, nonresponse error, measurement error, and processing error. Their definition of specification error, "when the concept implied by the survey question and the concept that should be measured in the survey differ" (38), is clearly akin to the notions of theoretical validity within psychometrics.

Biemer and Lyberg also are the first to attempt an integration of notions of process quality with total survey error. Process quality, continuous quality improvement, and a host of related notions arose in production environments spearheaded by Deming, Juran, Taguchi, and scores of others. Although some of these ideas have been oversold, oversimplified, or even misunderstood, the notions have been very useful in developing new approaches in survey work, such as the use of paradata (Couper 1998), methods to distinguish between common and special cause process variation using control charts (Morganstein

and Marker 1997), responsive design (Groves and Heeringa 2006), and minimizing waste in survey processes.

The book offers examples of situations in Industry and Occupation coding where continuous quality improvement efforts decrease the total survey error. The import of this discussion in the book is the raising, for the first time within a survey error format, the notion that "fitness for use" may be the most inclusive definition of quality. The authors offer no set of measurement approaches for "fitness for use" and follow with a more traditional treatment of survey errors. The reason that fitness for use is difficult to measure is that the notion encompasses not only the total survey error but also the qualitative nonstatistical dimensions mentioned earlier. Also, "fitness for use" can vary within surveys since most of the surveys are multipurpose and multiestimate. "Fitness for use" is a notion invented by Juran (1951) for quality in industry, while Deming used both "usefulness" and "fitness for purpose" for survey work. Also, Mahalanobis (1956) stated that "statistics must have purpose." This slight difference perhaps reflects the role of the customer in these two camps during the 1940s and 1950s.

The 2004 text *Survey Methodology* is organized around a total survey error framework, with an attempt to link the steps of survey design, collection, and estimation into the error sources. Figure 3 is a slight adaptation of the figure used in that text, which contains a term for the gap between the concept and the measure, labeled "validity," borrowing from psychometric notions of true score theory. In the original text, the figure is called "Survey lifecycle from a quality perspective." Chapters of the book link theory illuminating the causes of various error sources to design options to reduce the errors and practical tools that survey researchers use to implement the design options. The book attempts to note that two separate inferential steps are required in surveys—the first

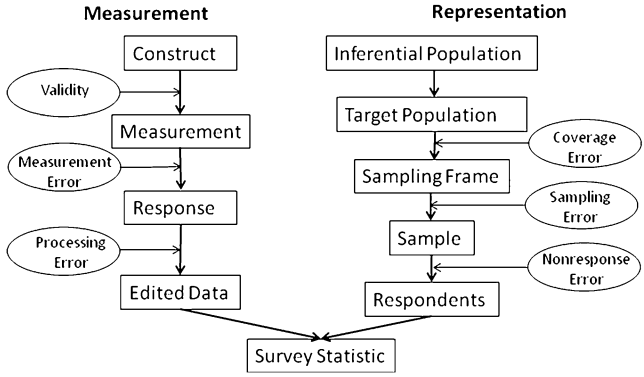

Figure 3. Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process (Groves et al. 2004).

inference is from the response to a question for a single respondent and the underlying construct of interest to the measurement. The second inference is from an estimate based on a set of respondents to the target population. The first of the inferential steps has been the focus of psychometric studies of measurement error, of simple response variance in surveys, and of correlated response variance in surveys. The second of these inferential steps focuses on coverage, nonresponse, and sampling error properties of sample-based statistics.

In his 2005 volume *The Total Survey Error Approach*, Herbert Weisberg presents a method for conducting survey research centered on the various types of errors that can occur. He also treats constraints such as ethical considerations and timeliness. He has a different take on processing errors, which are classified as Survey Administrative Issues in his typology of survey errors.

At this writing, the development of typologies continues. An alternative listing of error sources is offered by Smith (2009), who extends the total survey error paradigm to also include international surveys whose main purpose is to compare countries. Among other things he therefore adds the source "comparison error." He also adds the "conditioning error" that is unique for multi-wave panel studies. He claims that any typology has to be as detailed and comprehensive as possible.

To summarize, the roots of "total survey error" are found in lists of problems facing surveys beyond those of sampling error. The term evolves to a nested taxonomy of concepts of error, with major divisions based on variance and bias on one hand, and errors of observation and nonobservation on the other. In addition to these typologies serving as checklists of possible errors and their estimation, it is important to find ways to eliminate their root causes.

KEY STATISTICAL DEVELOPMENTS IN TOTAL SURVEY ERROR ESTIMATION

The statistical developments that led to practical estimators of survey error properties begin with some key observations about measurement. A notable stage in the evolution of total survey error is the initial raising of the issue of the existence of true values, to which biases might be linked. Hansen et al.'s (1951) paper published in *JASA* stands as the first extended statement of the concept. In this piece, the authors carefully present the differences between an "estimate" and the "value estimated." They lay out the three criteria of the notion of "true value," the first two being critical:

1. The true value must be *uniquely* defined.
2. The true value must be defined in such a manner that the purposes of the survey are met.

The final criterion is desirable, but not critical:

3. Where it is possible to do so consistently with the first two criteria, the true value should be defined in terms of operations, which can actually be carried through.

The second of the criteria, in many ways, is the equivalent to the notion of construct validity in the psychometrics literature (Lord and Novick 1968). Hansen et al. (1951) use the example that if one were interested in a child's intelligence, we would not define the true value as the score a teacher would assign on a particular day, but rather a more permanent attribute separated from the measurement itself. Hansen et al. acknowledge the complications of the notion of true value for some attributes. "What, for example, is a person's 'true intelligence,' 'true attitude toward revision of the Taft-Hartley Act,' or 'true brand preference for cigarettes?'" (p. 151). Deming (1960), on the other hand, claimed that true values do not exist, but that the value obtained from a realization of a "preferred survey technique" can be used as a proxy. This is in line with more recent discussions on "gold standards."

Since response errors were specific to measurements, the next key component was "essential survey conditions." This notion was key to what the appropriate universe of possible response values was. The universe was seen as conditioned on various properties of the measurement, including the subject of inquiry, mode of data collection, question structure, timing of the survey, sponsorship of the survey, interviewer corps, interviewing style, and recruitment protocol. "The expected value of the response errors, and the random component of variation around that expected value, may be regarded as determined by the essential survey conditions." Hansen et al. acknowledged that separating essential from nonessential survey conditions is difficult in practice, but that improvement of survey design requires acting on the essential ones.

The work of Mahalanobis (1946) and Hansen, Hurwitz, and Bershad (1961) using interpenetrated sample assignments to data production units (interviewers, coders, editors, keyers, supervisors, and crew leaders) was another notable development, which had long-run impact. The discovery that these production units might generate correlated response variances that in turn might result in effective sample sizes much below the intended ones was very important and led to fundamental changes in the 1960 U.S. Census data collection operations, where large parts of the enumeration were replaced by self-enumeration. To this day, this important finding is not fully known or appreciated in the survey community. Hansen et al. defined the mean squared error (MSE) of an estimate as the sum of a number of variance components and a squared bias term. In other words, MSE of an estimate could be decomposed as sampling variance + response variance + covariance of the response and sampling deviations + the squared bias. The response variance component could in turn be decomposed into simple response variance and correlated response variance. This is the well-known U.S. Bureau of the Census (which was the agency's name before the switch to the U.S. Census Bureau) survey model.

These developments coming from the inventors of key features of statistical sampling theory had impact. The "U.S. Bureau of the Census survey model" of measurement error variance is often mistaken as what "nonsampling error" means. However, it was a very limited model, ignoring coverage and nonresponse errors of all types. Further, the model was heavily constrained by what are clearly untenable assumptions:

1. The correlated response variance component is a function of the intraclass correlation coefficient among the response deviations for a survey, δ, and the sample size. In interview surveys, it was assumed that δ is independent of the workload of the interviewer (yet evidence abounds that interviewers alter their measurement error properties as a function of experience; see Rustemeyer 1977; Gfroerer, Eyerman, and Chromy 2002).
2. There was no acknowledgment that the attribute, y, being measured might be subject to inherent conceptual difficulties for some population elements (i.e., the construct validity of question/item might vary over subgroups).

Fellegi's (1964) article "Response Variance and Its Estimation" clarified some of these issues and identified design implications of the relaxing of these assumptions. Fellegi developed the variance of a sample mean as a function of sampling deviations and response deviations and their related covariances. These are the covariance terms, many of which Hansen, Hurwitz, and Bershad (1961) assume are zero, which were part of Fellegi's extension of the model:

1. correlation of response deviations obtained by different enumerators (e.g., arising from specific features of the implementation of training procedures)
2. correlation of sampling and response deviations *within enumerators* (e.g., tendency for enumerators to produce different deviations for elderly respondents than young respondents)

Fellegi notes that both interpenetration and reinterview designs are required to estimate these terms.

Hansen, Hurwitz, and Pritzker (1964) (see a fine historical review of this in Biemer 2004) developed a reinterview approach for estimating simple response variance by reasking questions of the respondent at a later point. They also discussed the effects of correlations between surveys (or trials) on the estimates of the simple response variance. They also provided results from experimental studies at the U.S. Bureau of the Census using reinterviews. While they were extending the Hansen, Hurwitz, and Bershad (1961) model, the authors communicated with Fellegi, which is evident from a note in Fellegi (1964, p. 1036). The fact that both interpenetration and reinterviews are necessary for the measurement of response variability components is discussed by Bailar and Dalenius (1969). They developed eight so-called basic study schemes for

the application of these methods and combinations of them in the estimation of response variance components. Although some of the schemes are impractical in the interview situation, they are all useful in handling response errors generated by coders, editors, supervisors, and crew leaders where the data collection organization can exert considerable control. What Hansen et al. did at this time was extend Neyman's randomization theory to include both randomization and nonsampling error theory. During a short period, this extension was in fact called "mixed error models" (Dalenius 1967).

The Tucson Measurement Error conference and monograph in 1990 (Biemer et al. 1991) was an important event in the evolution of the total survey error perspective because it attempted to bring together psychometric notions of measurement error and traditional survey notions of measurement error. Out of this conference came the Biemer and Stokes (1991) monograph chapter that took the reader from the simple true score model of psychometrics:

$$y_j = \mu_j + \varepsilon_j,$$

where y_i is the response to the question corresponding to construct μ_j on the jth person who produces error ε_j, to much more complex models involving interviewer variance and memory errors. The monograph also contains one of the first presentations of mixed structural and measurement error variance models using covariance analysis. Saris and Andrews (1991) present LISREL-based estimates of measurement error models, attempting to measure error influences of various mode and question structure properties, on structural models related latent variables to observables:

$$y_i = a_i + b_i F + g_i M_i + U_i + \varepsilon_i,$$

where y_i is the response for the ith item, measuring the latent variable F, using a method that is subject to the common component M_i and a joint influence of the trait and the method, U_i.

By creating multiple indicators measuring the same latent variable and representing a set of different methods, estimates of the coefficients in the model (which are estimates of validity, given the measurement model) can be extracted.

To summarize this developmental path, almost all of the statistical developments were focused on variance properties. The variance properties most often examined were measurement error variances. The most common tools to extract the estimates were reinterviews to provide multiple measures under the same essential survey conditions, interpenetration of interviewers to obtain correlated response variance components, and multiple indicators representing different methods, which provide validity estimates and correlated method variance. Admittedly, bias components also were studied at the time, but the statistical methods to do so were relatively straightforward. The general idea was that an

estimate is compared to an external source containing information on the same or part of the objects for which the estimate is sought. It is assumed that this external source is the result of a preferred procedure or is of "gold standard" caliber. It might be an existing record or a specific evaluation study. The general bias estimation technique is to calculate the difference between the survey estimate and the value obtained from the preferred procedure. Examples of evaluation studies at the time include Morgenstern (1963), Hurley, Jabine, and Larson (1962), and U.S. Bureau of the Census (1965).

SUMMARY OF THE HISTORICAL DEVELOPMENT OF TOTAL SURVEY ERROR

The enumeration of error sources in surveys began with a caution to users of surveys to avoid a singular focus on sampling error sources as the only important quality components. In short, Deming cautioned that sampling errors were only one of the quality criteria of survey statistics. The technical treatments of error sources focus most on variance components, and variance components on the measurement error side. One prominent discussion concerns inflation of measurement error variances because of shared response deviation tendencies among respondents interviewed by the same interviewer. Forsman (1989) provides a more detailed discussion of the early survey models and their use in survey-quality work.

Weaknesses of the Total Survey Error Framework

Having briefly reviewed the history of the notion of total survey error within the field of survey methodology, the remainder of the article critiques the development. We first identify the weaknesses of the concept, as now constituted; then we review its value.

EXCLUSIONS OF KEY QUALITY CONCEPTS

"Total Survey Error" is not the only way to think about information quality. Both statisticians and nonstatisticians have contributed separate notions of quality of information. Figure 4 presents a conceptual map constructed to aid the selection of quality criteria for the U.S. Key National Indicators Initiative, which aims to produce a continuously updated set of statistical indicators on the economy, the people, and the environment of the United States (Morillon 2005). The map incorporates many of the terms that are used in discussions among official statisticians throughout the world on data quality.

The figure intends to describe the quality of a statistical indicator (e.g., the unemployment rate, the median family income, the number of teenage women who gave birth in the last year, the consumer confidence index). The map uses the visual metaphor of a tree, with four principal branches: credibility, relevance, estimator quality, and data quality. Most of the statistical commentary

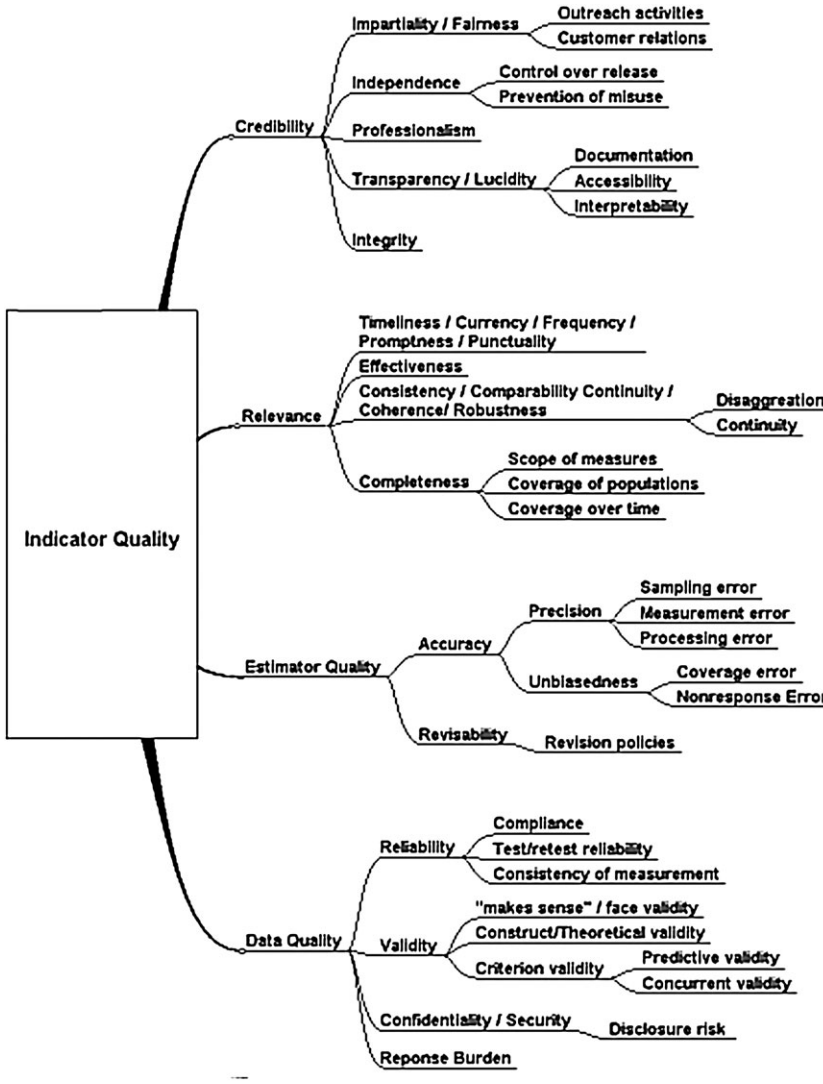

Figure 4. Draft Set of Indicator Quality Terms for Use by the Key National Indicators Initiative.

on total survey error would place the concept somewhere on the branch of estimator quality. (Some of the measurement error commentary would clearly be placed within the data quality branch, for example, that of psychometric notions of measurement error [Saris, van Wijk, and Scherpenzeel 1998; Biemer and Stokes 1991]).

Thus, it is appropriate to ask what is missing in the total survey error construction. Clearly, the user orientation to quality is to a large extent absent. Admittedly, there is an abundance of quality frameworks used in official statistics, mainly in some national statistical agencies in Europe and Oceania, Statistics Canada, and by organizations disseminating statistics from a group of countries, most notably Eurostat, UN, IMF, and OECD. These frameworks form the basis of quality reporting to their users, but as mentioned, we do not know much about how the bulk of users perceive information on survey errors and other nonstatistical dimensions of data quality. When it comes to this specific framework development, it seems as if producers of statistics have tended to put on hypothetical user hats and tried to imagine what they would like to know had they been users of statistics.

The literature on notions such as relevance and credibility is sparse. What is meant by those terms? Relevance of a statistic reflects the degree to which it meets the real needs of users. It is concerned with whether the available information sheds light on the issues that are important to users (Dalenius 1985). A similar thought is that "assessing relevance is subjective and depends upon the varying needs of users. . ." (Statistics Canada 2003). Dalenius thought that when errors are large and obvious, many users abstain from using the statistics, or if the statistics are irrelevant in the first place, errors do not matter. If relevance is viewed that way, then it is closely related to accuracy. Inaccurate statistics are not relevant, but irrelevant statistics might be accurate. Credibility or trust refers "to the confidence that users place in those products based simply on their image of the data producer, i.e., the brand image" (OECD 2003). Trust in the service provider is how most users judge the quality of statistical products (Trewin 2001), and trust has indeed been an issue for many organizations. The issue of trust can take many guises. The statistics might not be accurate, or it might be perceived as being presented in ways that are far from objective.

The conceptual map notes that a statistic may have desirable total survey error properties and still be assigned low quality by a user. The user judges the statistic to be of low quality because the credibility of the source is low (the source is seen as desiring the statistic to be, say, very high or very low, and therefore of low relevance to the user's own purposes). In this mythical case, the statistical properties of the estimate are quite desirable but the user-quality properties are quite undesirable. It is partly because of this reasoning that government statistical agencies place a high value on their independence from any political ideology dominant at the moment in the country. In contrast to this more inclusive framework, total survey error assumes a completely satisfactory correspondence between the error-free notion of the statistic and the purposes of the user.

What is the importance of the added notions of quality for survey methodology? First, they are expansive relative to the use of the statistics, not the estimates themselves. "Timeliness," "Relevance," "Credibility," "Accessibility,"

"Coherence," and other terms speak to how a particular use of a survey statistic matches the characteristics of the estimator.

Second, figure 4 differs from the total survey error framework in that it contains notions that are not easily quantifiable. Indeed, some are inherently unobservable properties of a statistic (e.g., "credibility" is an attribute brought by a user faced with a statistic). This takes them out of the natural domain of statistics.

Third, the framework, by adding other dimensions, forces attention to the early point of Deming (1944) already mentioned above, namely that a survey might be useful despite error problems. Total survey error does not have, within the framework, any notions of how to decide threshold acceptability of any of the levels of error. In short, loss functions regarding the quality of the decisions based on the estimates are rarely even conceptualized in optimal designs.

Fourth, any set of dimensions will contain pairs of dimensions that are in conflict with each other. For instance, nonresponse rate reduction is supposed to increase accuracy but takes time and will therefore affect timeliness negatively. Tradeoffs between quantitative and unmeasurable or qualitative dimensions are difficult to manage. From a design perspective, some of the nonstatistical dimensions could be viewed as fixed design constraints with a certain budget set aside to accommodate them. What is left is used to maximize accuracy.

LACK OF MEASURABLE COMPONENTS

The great disappointment regarding the total survey error perspective is that it has not led to *routine* fuller measurement of the statistical error properties of survey statistics. While official statisticians and much of the social science field have accepted the probability sampling paradigm and routinely provide estimates of sampling variance, there is little evidence that the current practice of surveys in the early 21st century measures anything more than sampling variance routinely.[1]

There are exceptions worth noting. The tendency for some continuing surveys to develop error or quality profiles is encouraging (Kalton, Winglee, and Jabine 1998; Kalton et al. 2000; Lynn 2003). These profiles contain the then-current understanding of statistical error properties of the key statistics produced by the survey. Through these quality profiles, surveys with rich methodological traditions produce well-documented sets of study results auxiliary to the publications of findings. None of the quality profiles have attempted full measurement of all known errors for a particular statistic.

1. Indeed, it is unfortunate that press releases of surveys sponsored by many commercial media remind readers of sampling error ("e.g., the margin of error for the survey is ±4 percentage points), while press releases of U.S. federal statistical agencies do not usually provide such warnings (see http://www.bls.gov/news.release/empsit.nr0.htm).

Also disappointing is the focus of statistical developments on the easily measurable response variance components or just on mere indicators of components and the lack of a partnership between the statistician and the subject-matter scientist in studies of biases in survey statistics.

Platek and Särndal (2001) and several colleagues discuss the role of survey methodology in statistics produced by national statistical agencies. Their main message is that the fact that we are not able to inform the user about all errors and the real uncertainty associated with estimates is very disturbing. On the other hand, the number of components of total survey error is not fixed, since new methods, technologies, and applications generate new or changed error structures.

INEFFECTIVE INFLUENCE TOWARD PROFESSIONAL STANDARDS FOR QUALITY MEASUREMENTS IN STATISTICAL PRESENTATIONS

Even without complete measurement of the error properties, one might have hoped by this point in history that a set of standard quality measures might have been endorsed by professional practice. It is obvious upon reflection that, without formalized loss and cost functions, the importance of knowledge of quality is itself a personal value of the user. Said differently, two questions are necessarily distinct: How important is quality of the information for a given use? How important is knowledge of the quality of information? Standards of official statistics practices have altered the values of users over time to demand sampling variance measures, and perhaps a few indicators of quality, such as nonresponse rates, although not various other errors within the total survey error perspective. Why is that? Measures of the different quality components have different costs, but quality itself has costs. For that reason, there is often a tension between achieving high quality and measuring quality (see Spencer 1985 for commentary on this issue). Also, most users are not in a position to comprehend quality components beyond sampling errors and nonresponse rates. Again, they might instead trust that other components are things the survey organization fully controls.

LARGE BURDENS FOR MEASUREMENT OF COMPONENTS

The limitation of the survey models of Hansen and his colleagues and Fellegi, besides being confined to measurement and sampling error, is that the estimation of their variance components requires interpenetrated designs and/or reinterviews. Both of these design features are costly. Thus, the designer faces the problem of how much of his/her budget should be spent on measuring quality versus on other things. Those studying interviewer variance components within the interpenetration design have come to the conclusion that the number of interviewers required to gain sufficient stability in estimates of the variance components (given the commonly low interviewer effects) is too large to be afforded (Groves and Magilavy 1986).

ASSUMPTIONS PATENTLY WRONG FOR LARGE CLASSES OF STATISTICS

Many of the error model assumptions are wrong most of the time. For example, the Kish (1962) linear model for interviewer effects assumes that the response deviations are random effects of the interviewer and respondent, uncorrelated to the true value for the respondent. However, for example, reporting of drug usage has been found to have an error structure highly related to the true value. Those who do not use drugs report "true" zeroes to frequency questions; those who are heavy users tend to commit underreporting errors. Villanueva (2001) notes that response deviation in self-reported body weight is a function of age, true weight, and body mass index. Later research has shown that self-reports of body weight also is a function of data collection mode (Beland and St-Pierre 2008). CATI respondents tended to report less body weight than self-respondents. Self-reports of income are known to be correlated with true values in a more complicated way, with lower-income persons overreporting and higher-income persons underreporting their incomes (Pedace and Bates 2001).

In short, there is growing evidence that measurement error models are highly variable by type of measurement (Alwin 2007; Biemer 2009). This diversity complicates the presentation of total survey error formulations. Indeed, it now seems clear that there are whole classes of measurement error models that need to be reformulated. Instead of concentrating on error models that have nice estimation properties, we should focus on error models that better reflect the phenomenon being modeled.

MISMATCH BETWEEN ERROR MODELS AND THEORETICAL CAUSAL MODELS OF ERROR

The existing survey models are specified as variance components models devoid of the causes of the error source itself. As methodological studies of measurement error have evolved, however, it is clear that the nature of the causes of measurement error is diverse. Missing in the history of the total survey error formulation is the partnership between scientists who study the causes of the behavior producing the statistical error and the statistical models used to describe them. This is a "two culture" problem that programs in survey methodology are actively attempting to eliminate, but have created too few products to demonstrate their success.

Why would causal models of statistical error be preferable? If we identify the causes of error, we can predict the error more accurately and even eliminate it. If we predict the error with the causes, then we can begin to act on the causes to either manipulate the cause (to reduce error) or measure the cause in order to alter the form of the estimation to reflect the cause. The increasing use of process data (paradata) in nonresponse rate reduction and postsurvey adjustment, motivated by theories of survey participation, is an example of this (Groves and Heeringa 2006). As a matter of fact, the use of process data to control processes

and to perform problem root-cause analysis is an alternative to more extensive evaluations of MSE components (Biemer and Lyberg 2003).

MISPLACED FOCUS ON DESCRIPTIVE STATISTICS

Hansen, Hurwitz, and Pritzker (1964); Hansen, Hurwitz, and Bershad (1961); Biemer and Stokes (1991); and Groves (1989) all describe error models for sample means and/or estimates of population totals. This is a tradition that flows out of the early developments in survey sampling, where descriptive statistics were the focus. These were valuable contributions. Yet survey data in our age are used only briefly for estimates of means and totals. Most analysts are interested more in subclass differences, order statistics, analytic statistics, and a whole host of model-based parameter estimates.

There has been relatively little development within survey statistics to examine error properties of such statistics. There is a literature on the effect of measurement errors on the univariate and multivariate analysis of survey data and how to compensate for the effects of such errors (Biemer and Trewin 1997). Also, it is known that coding errors can have an effect on the estimation of flows between occupations and when identifying subgroups for secondary analyses (Biemer and Lyberg 2003). It is, however, probably not obvious to scientific analysts of survey data how to embrace the total survey error paradigm.

FAILURE TO INTEGRATE ERROR MODELS FROM OTHER FIELDS

The isolation of survey statisticians and methodologists from the mainstream of social statistics has, in our opinion, retarded the importation of model-based approaches to many of the error components in the total survey error format. There are notable exceptions—the use of structural equation modeling (Andrews 1984; Saris and Andrews 1991), the use of multivariate models to unlock the essential survey conditions affecting response variance estimates (O'Muircheartaigh 1991), the use of hierarchical linear models to study interviewer variance (O'Muircheartaigh and Campanelli 1998; Hox 1994; Japec 2005); and the use of latent class models to understand response error components (Biemer 2010). These are enriching additions to the survey error literature. The field needs to integrate them more fully into the total survey error format and test the model applicability to wider sets of individual errors.

Strengths of the Total Survey Error Perspective

THE VALUE OF DECOMPOSITION OF ERRORS

Clearly, one of the strengths of the total survey error perspective is that it recognizes distinct ways that survey statistics can depart from some target parameter. It thus has the potential of protecting against various errors of inference.

This achievement was, indeed, the original intent of Deming's (1944) article. It worked.

What is the evidence that it has worked? Even media polls have methodological notes that specify that the survey results are affected by "question effects and other nonsampling errors." By isolating measurement error sources, the field has learned how to construct better questionnaires. By separating the types of nonresponse, we are learning to attack noncontact and refusal separately, given their different covariances with many survey variables. By separating mode effects from respondent effects, we have learned the powerful force of self-administration to reduce social desirability effects in reporting sensitive attributes. By separating coverage and nonresponse errors, we have begun to focus on the problematic nature of the frame unit of the household.

All of these advances are typical of science, when conceptual structures dominate a field. Careful decomposition of phenomena helps us understand how the components operate as a causal system. Taxonomies are important tools to organize thought and research. But, as pointed out by Deming (1944), ". . .the thirteen factors referred to are not always distinguishable and there are other ways of classifying them. . ."

SEPARATION OF PHENOMENA AFFECTING STATISTICS IN VARYING WAYS

The total survey error format forces attention to both variance and bias terms. This is not a trivial value of the framework. Most statistical attention to surveys is on the variance terms—largely, we suspect, because that is where statistical estimation tools are best found. Biases deserve special attention because

a. Their effect on inference varies as a function of the statistic; and
b. They are studied most often through the use of gold-standard measurements, not internal replication common to error variance.

The total survey error framework also permits useful focus on errors of observation versus errors of nonobservation. This too is useful. Errors of observation include those of measurement arising from the mode of data collection, interviewers, measurement instrumentation, and respondents themselves. Errors of nonobservation include coverage, nonresponse, and sampling errors. Noncoverage and nonresponse errors, despite the wonderful work in imputation and adjustment, are largely ignored by most analysts of data. But there are different possible reasons for this state of affairs. The analyst's motives might not always be in line with the decision-maker's needs, groups of methodologists tend to downplay the importance of other methodology groups' expertise, and there is an element of fashion in what is considered important current research that has an effect on what is being funded.

CONCEPTUAL FRAMEWORK OF THE FIELD OF SURVEY METHODOLOGY

The framework has facilitated the interdisciplinary nature of the field, permitting researchers from different disciplines to engage in discourse. For example, where sampling texts attend to variable error properties almost exclusively, the framework encourages attention to biases from various sources and thus the tradeoff of biases. The framework enriches the study of sampling error to attend to bias in variance estimators in the presence of nonresponse and measurement errors.

The framework seems undeniably to be the foundation of the field of survey methodology, shaping the discourse of the field. In graduate programs of survey methodology, it is the major identification of what subareas students are pursuing.

The framework offers an organizing principle for classification of the survey methodological literature. It permits the reader to critique that literature with regard to the limitations of its conclusions about various error sources. It facilitates the classification of the literature into those scientific studies of cause in error sources (e.g., studies of the causes of survey nonresponse), from those focused mainly on measurement of the extent of an error source, to those that study mainly the reduction of error through choice of different essential survey conditions.

By forcing attention to multiple error sources at once, it has laid out the unexplored research agenda in survey methodology. For example, while most modern conceptions of total survey error contain "processing error" as a component, that field is too rarely included in models of survey error, although Hansen, Hurwitz, and Pritzker (1964) explicitly incorporate coding as one of the response variance estimation applications of their model. Much of the early work in the field attempted to apply the same components of variance models to the coding step (e.g., Jabine and Tepping 1973; Kalton and Stowell 1979; Lyberg 1981). The same decomposition of the response variance into simple and correlated variance could be performed for coding as it was for interviewing, but with a simpler study situation. Current work attempts to blend computer software approaches to editing, with attention to blending editing and imputation schemes, with attention to efficiency and internal consistency as a quality criterion (Winkler 1999). Some attempts at replicated use of alternative editing systems on the same corpus of data do exist (Granquist 1997; Thompson 2002). Only by juxtaposing processing error next to other survey errors can the scientist observe the lacunae in the research literature and focus attention on those areas that need work.

A Vision of the Future

This last section of the article contains thoughts about the logical next steps in the development and enrichment of the total survey error framework.

STATUS OF SURVEYS AT THE BEGINNING OF THE MILLENNIUM

Surveys as tools of measurement and description of large human and establishment populations are always facing changing prospects because of three forces:

a. Researchers are continuously expanding the domains of knowledge they seek to study with the method (e.g., linkages between physiological and psychological health);
b. The populations being studied change their reaction to the request for measurement; and
c. The world external to surveys continuously offers new prospects for measurement (the Web, transaction data from administrative systems, digitized video images).

These three forces will shape the ongoing elaboration of the total survey error paradigm.

INTEGRATING CAUSAL MODELS OF SURVEY ERRORS

The early model of Hansen, Hurwitz, and Bershad (1961) and related developments (e.g., Stock and Hochstim 1951; Kish 1962) are variance components models, measuring the portion of instability in an estimate due to some feature of the essential survey conditions (e.g., interviewers or coders). The models themselves do not offer insights into what makes for large or small components. There is thus absent in the literature the questions of "why" that are typically asked by scientists. For instance, what question structure, attributes of words within the question, response categories, and response tasks produce higher interviewer variance? What types of respondents are most susceptible to interviewer variance? What different modes of data collection are most susceptible to interviewer variance? In short, what are the mechanisms within the measurement process that produce different levels of interviewer variance?

There is a large and complex research literature that does examine how interviewers affect survey results. For example, different training protocols do appear to affect interviewer variance (Fowler and Mangione 1990), differential interviewer probing produces different response distributions (Mangione, Fowler, and Louis 1992), and interviewers consistently monitored in centralized telephone facilities produce lower variability (Groves and Magilavy 1986).

Much of the insights from cognitive psychology into survey measurement properties could themselves specify a set of response error models. Some actually do; the measurement bias model associated with retrospective reporting of events (Sudman and Bradburn 1973) notes an exponential decay in the probability of reporting:

$$p_t = \beta_0 e^{\beta_1 t + \varepsilon},$$

where p_t is the probability that an event occurring at time t will be reported, β_0 is the probability of reporting at the time of the event itself, and β_t is the rate of decay in reporting over time.

As noted earlier (Groves 1998), other findings from that literature are not usually conceptualized as parameterized error models but easily could be. For example, to build on the Huttenlocher, Hedges, and Bradburn (1990) results, one could import the experimental findings of Smith and Jobe (1994), who note a two-part influence of response error to questions on food consumption. They posit effects of generic knowledge and episodic memory affecting reports. The generic knowledge produces an anchoring effect of a base rate of consuming a particular food, and episodic memories of events modify the base rate. It is then possible to represent the reporting bias as the difference between the estimate and the actual number of episodes remembered using the reporting decay function. The question is how the terms of such a model would be estimated. If, for instance, replicate subsamples were given questions that varied the reference period length at random, the form of the decay function might be possible to measure.

To sum, one research agenda for the future in total survey error is the specification of survey error models arising from social and behavioral science findings in survey behavior, or just findings that could have a bearing on survey behavior. One example of the latter is the use of findings regarding distributed cognition that could be used to illuminate response processes in establishment surveys (Lorenc 2006; Hutchins 1995).

INTERPLAY OF DIFFERENT ERROR SOURCES

A notable omission in the total survey error models is a set of parameters that describe the relationship between two error magnitudes or among multiple error sources. It is typically the case that these terms are assumed to be absent. For example, consider the covariance between nonresponse bias and measurement bias terms. Cannell and Fowler (1963) show that respondents with lower response propensities tended to have a higher likelihood of measurement biases. That is, there is a positive covariance between the likelihood of nonresponse and the likelihood of measurement error. Also, there is at least anecdotal evidence that interviewers that achieve high response rates tend to have issues with measurement biases.

The absence of this particular term linking nonresponse and measurement error, for example, is disturbing, given the finding that there is evidence that interviewer variance components estimated merely by respondent attributes (i.e., subject to nonresponse biases) are generally larger than those controlling on key attributes of respondents known to be subject to interviewer variance in response rates. That is, the traditional estimates of interviewer variance are combined estimates of nonresponse and response error variance.

These are not trivial omissions, especially when one brings cost constraints into the design picture. It is difficult to overstate how much of current budgets for household survey data collection are devoted to reducing nonresponse rates (through repeated calls to reduce noncontacts, reassignment of interviewers for non-English speakers, and refusal conversion efforts). Reduction of nonresponse rates, if productive of higher measurement error components, may be counterproductive for total survey error. Even mild tendencies in that direction may suggest lower optimal response rates to minimize mean squared error per unit cost. This suggestion is in conflict with the view that high response rates are an indication of agency competence and user trust in the agency. These kinds of tradeoffs might be hard to sell even though they can be proven efficient.

INTERPLAY OF BIAS AND VARIANCE COMPONENTS

The invention of new reduction tools for survey bias (through more effective question structure and self-administered modes) means that it is more important for us to understand how variance and bias terms relate to one another. For example, if conversational interviewing succeeds in reducing response biases due to comprehension errors (Schober and Conrad 1997), what will happen to the correlated response variance component? If interviewer training can show reduced nonresponse bias and reduced variance in response rates across interviewers, is the nonresponse error variance of a statistic reduced (Groves and McGonagle 2001)? If self-administration reduces underreporting bias for sensitive attributes, does it also affect simple response variance? There are many other examples of questions like these.

To make progress on such questions, studies that jointly examine biases and variances need to be mounted. Such studies require a deliberate effort on the part of the survey methodological community and are currently encouraged by the International Total Survey Error Workshop (http://niss.org/itsew).

THE DEVELOPMENT OF DIAGNOSTIC TOOLS TO STUDY MODEL ERROR IN MODEL-BASED ESTIMATES OF SURVEY ERROR COMPONENTS

The social sciences have developed standards for model assessment that demand that the researcher examines assumptions of the model, contrast conclusions based on multiple model specification, and entertain alternative causal logic. Much of the survey error estimation asserts a single model, within a given set of assumptions, and derives all conclusions about the error properties from that model.

Work needs to begin to address alternative model specifications (and the designs that go with them) and demonstrations of the impact of violations of the assumptions of the models on conclusions about survey quality. Such a line of development is the logical next step after error models have been specified.

ASSOCIATING THE VALUE OF QUALITY MEASUREMENT WITH THE VALUE OF QUALITY ITSELF

Outside the total survey error paradigm lies an unanswered question: Given that measuring quality and increasing quality both cost money, how much of the budget should be allocated to each? In short, in addition to the question "What price does quality justify?" we have "What price do measures of quality justify?" and "How much should we spend on increasing quality and how much on measuring quality?" Alternative questions are "How much should be spent on prevention, control, and improvement, or on pretesting, monitoring, and evaluation?" Many survey managers are worried about error prevention costs. It is therefore important to also look at failure and appraisal costs, figures that are seldom secured in survey organizations. Costs of rework when inspection results so indicate, rework as a result of large errors in databases and published material, handling client complaints, production of statistics with low relevance, and costs for running inefficient processes have a tendency to remain invisible if managers do not look for them, but instead they appear as "lack of financial resources." Some of these failure and appraisal costs would disappear if processes were less error prone (Lyberg, Japec, and Biemer 1998). In any case, these allocation issues are imperative to survey quality.

These are not completely new issues for survey statistics. For years, survey samplers have argued about the merits of deeper stratification permitted in a systematic sample on an ordered list, yielding biased estimates of sampling variance, versus a paired selection design from fixed strata, yielding an unbiased estimator for the sampling variance. Most practical samplers forgo the unbiased estimate of the sampling variance (an error measurement) for the "assumed" lower magnitude of the true sampling variance.

Spencer (1985) notes that many approaches to the issue assert that more "important" decisions using statistical information demand more complete measurement of the quality of the statistics. This perspective creates a problem for designers who are unaware of the myriad uses of a statistical data collection prior to its inception. However, it is no more perplexing than the job of multipurpose optimization that has faced designers for years.

INTEGRATING OTHER NOTIONS OF QUALITY INTO THE TOTAL SURVEY ERROR PARADIGM

Because statistics are of little importance without their use being specified, "fitness for use" is of growing importance as a quality concept. The integration of the notion of fitness for use might be partially related to notions of "construct validity" in psychometrics. It is relatively common for national statistical agencies to refer to their quality frameworks as a means to achieve fitness for use. When this is the case, there is a risk that the fitness-for-use concept is watered down. The Australian Bureau of Statistics has made a concerted effort to define and help users interpret the concept. We believe that this organization is at the

forefront among agencies when it comes to operationalizing the concept. It even has a User Centered Design Unit (Tam and Kraayenbrink 2007) to coordinate their service deliveries.

Communication about quality is an important judgment in "fitness for use." This is especially true when one statistic may be used with very different purposes. Then, tradeoffs of bias and variance can become important, as well as tradeoffs of coverage and nonresponse versus various types of measurement errors. Sample surveys have won over many societies, and statistical information is the standard metric of public discourse. Given this, it is appropriate for the field to bring into the total survey error perspective the notion of use-specific error and to address how best to conceptualize it and, if possible, invent proxy measures of it.

HOW A MULTI-MODE, MULTI-FRAME, MULTI-PHASE WORLD MAY HELP US

For some time, many survey methodologists have been predicting a movement away from single-mode, single-frame survey designs to multiple-frame, multiple-mode designs (Lepkowski and Groves 1986; Groves 1987; de Leeuw 2005; Dillman and Tarnai 1988; Dillman, Smyth, and Christian 2009). In fairness, it has already occurred in many domains of surveys—panel surveys are mixing telephone and face-to-face surveys; ACASI use embedded in face-to-face surveys blends self-administered and interviewer-administered data collection; and multi-phase surveys mix modes to raise response rates (e.g., the American Community Survey).

How is this relevant to total survey error? Mixed-mode, multi-frame, multi-phase designs alter the essential survey conditions of a survey. Multiple essential survey conditions offer built-in contrasts that can be informative about error properties of survey statistics. It is likely to be true that this variation in essential survey conditions will not always be coextensive with separate replicate samples. Hence, it is ripe terrain for the model-builders among us, to specify both designs and model-based error estimates of value to the researcher.

Summary

The total survey error paradigm is the conceptual foundation of the field of survey methodology. Its roots lie in cautions by the founders of sample surveys that the quality properties of survey statistics are functions of essential survey conditions that are independent of the sample design.

The paradigm has been more successful as an intellectual framework than a unified statistical model of error properties of survey statistics. More importation of modeling perspectives from other disciplines could enrich the paradigm.

The current weaknesses in the status of research on total survey error are that key quality concepts are not included (notably those that take a user perspective), that quantitative measurement of many of the components lags their

conceptual richness, that the paradigm has not led to enriched error measurement in practical surveys, that there are large burdens to the measurement of some components, that assumptions required for some estimators of error terms are frequently not true, that there is a mismatch between the existing error models and theoretical causal models of the error mechanisms, that there is a misplaced focus on descriptive statistics, and that there is a failure to integrate error models developed in other fields.

The strengths of the total survey error framework are the explicit attention to the decomposition of errors, the separation of phenomena affecting statistics in various ways, and its success in forming the conceptual basis of the field of survey methodology, pointing the direction for new research.

The future research agenda in total survey error should include the union of causal models of error propagation and error measurement, the study of the interplay of various different error sources, the interplay of bias and variance components within a single error source, the development of methods for assessing risks associated with single error sources, the role of standards for measuring and reporting sources of error, the development of diagnostic tools to study model error, clarification of the role of quality measurement within efforts to increase quality, integrating other notions of quality into the framework, exploiting mixed-mode designs as vehicles to measure various error components, and the role of the total survey error framework in a survey organization's quality management system.

References

Alwin, Duane. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. New York: Wiley.

Anderson, Ronald, Judith Kasper, and Martin Frankel. 1979. *Total Survey Error: Applications to Improve Health Surveys*. San Francisco, CA: Jossey-Bass.

Andrews, Frank. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach." *Public Opinion Quarterly* 48:409–42.

Bailar, Barbara, and Tore Dalenius. 1969. "Estimating the Response Variance Components of the U.S. Bureau of the Census Survey Model." *Sankhya, Series B* 31:341–60.

Beland, Yves, and Martin St-Pierre. 2008. "Mode Effects in the Canadian Community Health Survey: A Comparison of CATI and CAPI." In *Advances in Telephone Survey Methodology,* eds. James Lepkowski, Clyde Tucker, Michael Brick, Edith de Leeuw, Lilli Japec, Paul Lavrakas, Michael Link, and Roberta Sangster. New York: Wiley, 297–314.

Biemer, Paul. 2004. "Simple Response Variance: Then and Now." *Journal of Official Statistics* 20:417–39.

———. 2009. "Measurement Error in Sample Surveys." In *Handbook of Statistics 29A,* eds. Danny Pfefferman and C.R. Rao. North-Holland, Chapter 12.

———. 2010. *Latent Class Analysis of Survey Error*. New York: Wiley.

Biemer, Paul, and Lars Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Biemer, Paul, Robert Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman, eds. 1991. *Measurement Errors in Surveys*. New York: Wiley.

Biemer, Paul, and Lynne Stokes. 1991. "Approaches to the Modeling of Measurement Error." In *Measurement Errors in Surveys,* eds. Paul Biemer, Robert Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman. New York: Wiley, 487–516.

Biemer, Paul, and Dennis Trewin. 1997. "A Review of Measurement Error Effects on the Analysis of Survey Data." In *Survey Measurement and Process Quality,* eds. Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: Wiley, 603–32.

Brackstone, Gordon. 1999. "Managing Data Quality in a Statistical Agency." *Survey Methodology* 25(2):139–49.

Cannell, Charles, and Floyd Fowler. 1963. "A Study of the Reporting of Visits to Doctors in the National Health Survey." Research Report, Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Cochran, William. 1953. *Sampling Techniques*. New York: Wiley.

———. 1968. "Errors of Measurement in Statistics." *Technometrics* 10:637–66.

Couper, Mick. 1998. "Measuring Survey Quality in a CASIC Environment." Paper presented at the Joint Statistical Meetings, American Statistical Association, Dallas, TX.

Dalenius, Tore. 1967. "Nonsampling Errors in Census and Sample Surveys." Report No. 5 in "Errors in Surveys," Stockholm University.

———. 1974. "Ends and Means of Total Survey Design" Report in "Errors in Surveys," Stockholm University.

———. 1985. "Relevant Official Statistics." *Journal of Official Statistics* 1(1):21–33.

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

Deming, Edwards. 1944. "On Errors in Surveys." *American Sociological Review* 9:359–69.

———. 1950. *Some Theory of Sampling*. New York: Wiley.

———. 1960. *Sample Design in Business Research*. New York: Wiley.

Dillman, Don, Jolene Smyth, and Leah Christian. 2009. *Internet, Mail, and Mixed-mode Surveys*. New York: Wiley.

Dillman, Don, and John Tarnai. 1988. "Administrative Issues in Mixed-mode Surveys." In *Telephone Survey Methodology,* eds. Robert Groves, Paul Biemer, Lars Lyberg, James Massey, William Nicholls II, and Joseph Waksberg. New York: Wiley.

Eurostat. 2000. "Assessment of the Quality in Statistics." Eurostat General/Standard Report, Luxembourg, April 4–5.

Fellegi, Ivan. 1964. "Response Variance and Its Estimation." *Journal of the American Statistical Association* 59:1016–41.

Felme, Sven, Lars Lyberg, and Lars Olsson. 1976. *Kvalitetsskydd av data* (Protecting Data Quality). Liber (in Swedish).

Forsman, Gösta. 1989. "Early Survey Models and Their Use in Survey Quality Work." *Journal of Official Statistics* 5:41–55.

Fowler, Floyd, and Thomas Mangione. 1990. *Standardized Survey Interviewing*. Newbury Park, CA: Sage Publications.

Gfroerer, Joseph, Joe Eyerman, and James Chromy, eds. 2002. *Redesigning an Ongoing Survey: Methodological Issues*. DHHS Publication No. SMA 03-3768. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Granquist, Leopold. 1997. "An Overview of Methods of Evaluating Data Editing Procedures." *Proceedings for the Conference of European Statisticians Statistical Standards and Studies, Section on Statistical Data Editing (Methods and Techniques).* United Nations Statistical Commission and Economic Commission for Europe, 112–23.

Groves, Robert. 1987. "Research on Survey Data Quality." *Public Opinion Quarterly* 51:156–72.

———. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

———. 1998. "Survey Error Models and Cognitive Theories of Response Behavior." In *Cognition and Survey Research,* eds. Monroe Sirken, Douglas Hermann, Susan Schechter, Norbert Schwarz, Judith Tanur, and Roger Tourangeau. New York: Wiley, 235–52.

Groves, Robert, Floyd Fowler, Mick Couper, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. New York: Wiley.

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

Groves, Robert, and Lou Magilavy. 1986. "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys." *Public Opinion Quarterly* 50:251–66.

Groves, Robert, and Katherine McGonagle 2001. "A Theory-guided Interviewer Training Protocol Regarding Survey Participation." *Journal of Official Statistics* 17:249–66.

Hansen, Morris, William Hurwitz, and Max Bershad 1961. "Measurement Errors in Censuses and Surveys." *Bulletin of the International Statistical Institute,* 32nd Session 38, Part 2:359–74.

Hansen, Morris, William Hurwitz, and William Madow. 1953. *Sample Survey Methods and Theory* Volumes I–II. New York: Wiley.

Hansen, Morris, William Hurwitz, Eli Marks, and Parker Mauldin. 1951. "Response Errors in Surveys." *Journal of the American Statistical Association* 46:147–90.

Hansen, Morris, William Hurwitz, and Leon Pritzker. 1964. "The Estimation and Interpretation of Gross Differences and Simple Response Variance." In *Contributions to Statistics,* ed. C. Rao. Oxford, UK: Pergamon Press, 111–36.

———. 1967. "Standardization of Procedures for the Evaluation of Data: Measurement Errors and Statistical Standards in the Bureau of the Census." Paper presented at the 36th Session of the International Statistical Institute.

Hert, Carol, and Sheila Denn. 2006. "Understanding the Role of Metadata in Finding and Using Statistical Information." Invited paper to the UN Work Session on Statistical Dissemination and Communication, Washington DC, September 12–14.

Hox, Joop. 1994. "Hierarchical Regression Models for Interviewer and Respondent Effects." *Sociological Methods & Research* 22:300–318.

Hurley, Ray, Thomas Jabine, and Don Larson. 1962. "Evaluation Studies of the 1959 Census of Agriculture." *American Statistical Association, Proceedings of the Social Statistics Section,* 91–95.

Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge: Massachusetts Institute of Technology.

Huttenlocher, Janellen, Larry Hedges, and Norman Bradburn. 1990. "Reports of Elapsed Time-bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology, Learning, Memory, and Cognition* 16:196–213.

Jabine, Thomas, and Ben Tepping. 1973. "Controlling the Quality of Occupation and Industry Coding." *Bulletin of the International Statistical Institute,* 360–92.

Japec, Lilli. 2005. *Quality Issues in Interview Surveys: Some Contributions*. PhD dissertation, Stockholm University.

Juran, Joseph. 1951. *Quality Control Handbook*. New York: McGraw-Hill.

Kalton, Graham, and Richard Stowell. 1979. "A Study of Coder Variability." *Journal of the Royal Statistical Society, Series C* 28:276–89.

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

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

Kish, Leslie. 1962. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association,* 92–115.

———. 1965. *Survey Sampling*. New York: Wiley.

Lepkowski, James, and Robert Groves. 1986. "A Mean Squared Error Model for Dual-frame, Mixed-mode Surveys." *Journal of the American Statistical Association* 81:930–37.

Lessler, Judith, and William Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: Wiley.

Lord, Frederic, and Melvin Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company.

Lorenc, Boris. 2006. *Modeling the Response Process in Establishment Surveys*. PhD dissertation, Stockholm University.

Lyberg, Lars. 1981. *Control of the Coding Operation in Statistical Investigations: Some Contributions*. PhD dissertation, Stockholm University.

Lyberg, Lars, Lilli Japec, and Paul Biemer. 1998. "Quality Improvement in Surveys: A Process Perspective." *American Statistical Association, Proceedings of the Survey Research Methods Section,* 23–31.

Lynn, Peter, ed. *2003. Quality Profile: British Household Panel Survey: Waves 1 to 10: 1991–2000.* Colchester, UK: Institute for Social and Economic Research.

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

———. 1956. Speeches on the Occasion of the 25th Anniversary of the Indian Statistical Institute.

Mangione, Thomas, Floyd Fowler, and Thomas Louis. 1992. "Question Characteristics and Interviewer Effects." *Journal of Official Statistics* 8:293–307.

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

Morgenstern, Oskar. 1963. *On the Accuracy of Economic Observations*. Princeton, NJ: Princeton University Press.

Morillon, Anne. 2005. Personal communication.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society,* 97:558–606.

OECD. 2003. *Quality Framework and Guidelines for Statistical Activities*. Version 2003/1, http://www.oecd.org/dataoecd/26/42/21688835.pdf.

O'Muircheartaigh, Colm. 1991. "Simple Response Variance: Estimation and Determinants." In *Measurement Error in Surveys,* ed. Paul Biemer, Robert Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman. New York: Wiley, 551–74.

O'Muircheartaigh, Colm, and Pamela Campanelli. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistical Society, Series A* 161:63–77.

Pedace, Roberto, and Nancy Bates. 2001. "Using Administrative Records to Assess Earnings Reporting Error in the Survey of Income and Program Participation." *Journal of Economic and Social Measurement* 26:173–92.

Platek, Richard, and Carl-Erik Särndal. 2001. "Can a Statistician Deliver?" *Journal of Official Statistics,* 17(1):1–20 and discussion 21–127.

Rustemeyer, Anitra. 1977. "Measuring Interviewer Performance in Mock Interviews." *Proceedings of the American Statistical Association, Social Statistics Section*. Alexandria, VA: American Statistical Association, 341–46.

Saris, Willem, and Frank Andrews. 1991. "Evaluation of Measurement Instruments Using a Structural Modeling Approach." In *Measurement Error in Surveys,* eds. Paul Biemer, Robert Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman. New York: Wiley, 575–97.

Saris, Willem, Theresia van Wijk, and Annette Scherpenzeel. 1998. "Validity and Reliability of Subjective Social Indicators: The Effect of Different Measures of Association." *Social Indicators Research* 45:173–99.

Schober, Michael, and Fred Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly* 61:576–602.

Smith, Albert, and Jared Jobe. 1994. "Validity of Reports of Long-term Dietary Memories: Data and a Model." In *Autobiographical Memory and the Validity of Retrospective Reports,* eds. Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag, 121–40.

Smith, Tom. 2009. "Refining the Total Survey Error Perspective." Memo: NORC/University of Chicago.

Spencer, Bruce. 1985. "Optimal Data Quality." *Journal of the American Statistical Association* 80:564–73.

Statistics Canada. 2003. *Quality Guidelines*, 4th ed. Ottawa: Statistics Canada.

Stock, J. Stevens, and Joseph Hochstim. 1951. "A Method of Measuring Interviewer Variability." *Public Opinion Quarterly* 15:322–34.

Sudman, Seymour, and Norman Bradburn. 1973. "Effects of Time and Memory Factors on Response in Surveys." *Journal of the American Statistical Association* 68:805–15.

Tam, Siu-Ming, and Regina Kraayenbrink. 2007. "Data Communication—Emerging International Trends and Practices of the Australian Bureau of Statistics." *Statistical Journal of the United Nations Economic Commission for Europe* 23(4):229–47.

Thompson, Katherine. 2002. "Evaluating the Effect of Two Editing Systems on Data Quality: Two Case Studies." UNECE Statistical Data Editing Work Session, Helsinki.

Trewin, Dennis. 2001. "Measuring Well-being: Frameworks for Australian Social Statistics." Memo, Australian Bureau of Statistics.

U.S. Bureau of the Census. 1965. "Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Employer Record Check." Series ER 60, No. 6, U.S. Government Printing Office, Washington, DC.

Villanueva, Elmer. 2001. "The Validity of Self-reported Weight in U.S. Adults: A Population Based Cross-sectional Study." *BMC Public Health* 1.

Weisberg, Herbert. 2005. *The Total Survey Error Approach*. Chicago, IL: University of Chicago Press.

Winkler, William. 1999. "State of Statistical Editing and Current Research Problems." In *Statistical Data Editing*. Rome, Italy: ISTAT 169–87.

Zarkovich, Slobodan. 1966. *Quality of Statistical Data*. Rome, Italy: Food and Agricultural Organization of the United Nations.