1.1 Introduction and Historical Backdrop

In this chapter, we discuss the concept of total survey error (TSE), how it originated and developed both as a mindset for survey researchers and as a criterion for designing surveys. The interest in TSE has fluctuated over the years. When Jerzy Neyman published the basic sampling theory and some of its associated sampling schemes in 1934 onward, it constituted the first building block of a theory and methodology for surveys. However, the idea that a sample could be used to represent an entire population was not new. The oldest known reference to estimating a finite population total on the basis of a sample dates back to 1000 BC and is found in the Indian epic Mahabharata (Hacking, 1975; Rao, 2005). Crude attempts at measuring parts of a population rather than the whole had been used in England and some other European countries quite extensively between 1650 and 1800. The methods on which these attempts were based were referred to as political arithmetic (Fienberg and Tanur, 2001), and they resembled ratio estimation using information of birth rates, family size, and average number of persons living in selected buildings and other observations. In 1895, at an International Statistical Institute meeting, Kiaer argued for developing a representative or partial investigation method.
The representative method aimed at creating a sample that would reflect the composition of the population of interest. This could be achieved by using balanced sampling through purposive selection or various forms of random sampling. During the period 1900–1920, the representative method was used extensively, at least in Russia and the U.S.A. In 1925, the International Statistical Institute released a report on various aspects of random sampling (Rao, 2005, 2013; Rao and Fuller, 2015).

The main consideration regarding sampling was likely monetary, given that it was resource-intensive and time-consuming to collect data from an entire population. Statistical information compiled using a representative sample was an enormous breakthrough. But it would be almost 40 years after Kiaer’s proposal before Neyman published his landmark paper from 1934 “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection.” At this time, there existed some earlier work by the Russian statistician Tschuprow (1923a, b) on stratified sampling and optimal allocation. It is not clear whether Neyman was aware of this work when he started to develop the sampling theory in the 1920s (Fienberg and Tanur, 1996) since he did not mention Tschuprow’s work when discussing optimal allocation. Neyman definitely had access to Ronald Fisher’s (1925) ideas on randomization (as opposed to various kinds of purposive selection) and their importance for the design and analysis of experiments, and also to Bowley’s (1926) work on stratified random sampling.

The sampling methods proposed by Neyman were soon implemented in agencies such as the Indian Statistical Institute and the U.S. Bureau of the Census (currently named the U.S. Census Bureau). Prasanta Mahalanobis, the founder of the Indian Statistical Institute, and Morris Hansen and colleagues at the U.S. Census Bureau, became the main proponents of scientific sampling in a number of surveys in the 1940s. The development was spurred on by Literary Digest’s disastrously inaccurate prediction in the 1936 U.S. presidential election poll that was based on a seriously deficient sampling frame. However, Neyman’s sampling theory did not take into account nonsampling errors and relied on the assumption that sampling was the only major error source that affected estimates of population parameters and associated calculations of confidence intervals or margins of error. However, Neyman and his peers understood that this was indeed an unrealistic assumption that might lead to understated margins of error. The effect of nonsampling errors on censuses was acknowledged and discussed in a German textbook on census methodology relatively early on (Zizek, 1921). The author discussed what he called control of...
In addition, Karl Pearson (1902) discussed observer errors much earlier than that. An early example of interviewer influence on survey response was the study on the consumption of hard liquor during the prohibition days in the U.S.A., where Rice (1929) showed that interviewers who were prohibitionists tended to obtain responses that mirrored their own views and that differed from those of respondents that were interviewed by other interviewers.

In 1944, Edwards Deming published the first typology of sources of error beyond sampling. He listed 13 factors that he believed might affect the utility of a survey. The main purpose of the typology was to demonstrate the need for directing efforts to all potential sources in the survey planning process while considering the resources available. This first typology included some error sources that are not frequently referenced today, such as bias of the auspices (i.e., the tendency to indicate a particular response because of the organization sponsoring the study). Others, to which more attention is currently given, such as coverage error, were not included, however. Even though Deming did not explicitly reference TSE, he emphasized the limitations of concentrating on a few error sources only and highlighted the need for theories of bias and variability based on accumulated experience.

Rapid development of the area followed shortly thereafter. Mahalanobis (1946) developed the method of interpenetration, which could be used to estimate the variability generated by interviewers and other data collectors. Another error source recognized early on was nonresponse. Hansen and Hurwitz (1946) published an article in the Journal of the American Statistical Association on follow-up sampling from the stratum of initial nonrespondents. While the basic assumption of 100% participation in a follow-up sample was understood not to be realistic, at the time, there were relatively small nonresponse rates, and it was possible to estimate, at least approximately, the characteristics of those in the nonresponse stratum.

Even though it is not explicitly stated, TSE has its roots in cautioning against sole attention focused on sampling error along with possibly one or two other error sources, rather than the entire scope of potential errors. In response, two lines of strategic development occurred. One strategy entailed the identification of specific error sources, coupled with an attempt to control them or at least minimize them. The other strategy entailed the development of the so-called survey error models, where the TSE was decomposed and the magnitude of different error components, and ultimately the combination of them (i.e., the TSE), could be estimated. The two strategies were intertwined in the sense that a survey model could be applied not only on the entire set of survey operations but also on a subset of specific survey operations.

1.2 Specific Error Sources and Their Control or Evaluation

Apart from that of Deming (1944), there are a number of typologies described in the survey literature. Examples include Kish (1965), Groves (1989), Biemer and Lyberg (2003), Groves et al. (2009), Smith (2011), and Pennell et al. (Chapter 9 in this volume). Some of them are explicitly labeled TSE, while others consist of listings of different types of errors; however, all are incomplete. In some cases, known error sources (as well as their interactions with other error...
sources) are simply omitted, and in other cases, all possible error sources are not known or the sources defy expression. For instance, new error structures have emerged when new data collection modes or new data sources, such as Big Data (see, e.g., Chapter 3 in this volume), have become popular—but the comprehension and articulation of the associated error structures have lagged in time.

Early on, the work toward the treatment of specific error sources followed two separate types of strategies: control and evaluation.

Related to the first strategy of control, one line of thinking was that statistical agencies were “data factories” that produced tables and limited analyses as their outputs. As such, they resembled an industrial assembly line. Therefore, the application of methods for industrial quality control (QC) was deemed suitable. Several statistical agencies adopted this approach for some of their operations, and the U.S. Census Bureau was clearly at the forefront. Most of these QCs were focused toward manual operations such as enumeration and interviewing, listing, coding, card punching, and editing, although it was also possible to use QC to check automatic operations such as scanning, which at the time was implemented through Film Optical Sensing Device for Input to Computers (FOSDIC). For the manual operations, the main control method was verification, where one operator’s work was checked by another operator. A long list of census methodologists including Morris Hansen, Bill Hurwitz, Eli Marks, Edwards Deming, Ross Eckler, Max Bershad, Leon Pritzker, Joe Waksberg, Herman Fasteau, and George Minton made very significant contributions to this QC development. Contributions included those of Deming et al. (1942), Hansen and Steinberg (1956), Hansen et al. (1962), and the U.S. Bureau of the Census (1965).

These QC schemes were adapted from their industrial applications, and therefore were called “administrative applications of statistical QC.” One example of this kind of scheme related to the coding of variables with respect to Industry and Occupation (Fasteau et al., 1964). During that era, a coder’s work was typically verified by one or more coders in a dependent or independent way. To protect the users of data, acceptance sampling schemes were applied. Under such schemes, coding cases were bundled together in lots and sample inspection took place. If the number of coding errors was below or equal to an acceptance number, the lot was accepted. However, if the number of coding errors exceeded the acceptance number, the lot underwent 100% inspection, after which a decision was made that a coder should either remain on sampling control or remain under total control until results improved. An added complication was the institution of a point system that was imposed on the coders. Under the point system, the coder was given an initial allotment of three points. When a favorable quality decision was made, the coder received one more point. Otherwise, he or she lost one point. When the accumulated point balance reached zero, remedial action was taken toward the coder either in the form of additional training or dismissal from the operation. To avoid excessive accumulation of points that might culminate during a long period and that might mask substandard coding, the accumulated score was adjusted after every 10th decision. If the accumulated score was above 3 after the 10th decision it was reduced to 3. If the accumulated score was 3, 2, or 1, the coders maintained their current score (Minton, 1970).
One element that was often lacking with this factory approach was productive feedback, because at the time, root cause analysis was not really seen as a priority and “rework” was the prescription. Acceptance sampling was later vigorously criticized by Deming (1986), who claimed that under such a system, continuous improvement could not be achieved.

During the next decade, these schemes became increasingly complicated and were eventually abandoned in place of automated systems (Minton, 1972). It should be mentioned, though, that coding errors could be and still remain to this day quite substantial. Even today, gross errors, the difference between a production code and a verification code, in the range of 10–20% are not unusual. In present day systems, coding is often performed by software, but the error source itself is still basically neglected in most statistical agencies (Groves and Lyberg, 2010). The contributing factors are, in part, due to lack of software upgrades and minimal control of residual manual coding.

Another source of nonsampling error that received a lot of attention over the years is unit nonresponse. In the 1950s and 1960s, nonresponse was seen as catastrophic in terms of the ability to ensure high quality of survey results. Even modest nonresponse rates could trigger very unrealistic reactions, where fears that all nonrespondents might have values different from the respondents were prevalent. For instance, in the measurement of the unemployment rate, if in the extreme, all nonrespondents are assumed to be either employed or unemployed, it would then be possible to create max–min intervals that produced a much exaggerated picture of the risk and impact of nonresponse (Dalenius, 1961). This rigid view was later replaced by adjustment methods (Kalton and Kaspryzk, 1986), and theories and methods for missing data (Rubin, 1976). In addition, monographs on nonresponse and missing data (Groves et al., 2002; Madow et al., 1983) were written, as were textbooks on specific treatments of nonresponse such as multiple imputation (Rubin, 1987), theories of survey participation (Groves and Couper, 1998), and nonresponse in the international European Social Survey (Stoop et al., 2010). Brick (2013) reviewed various adjustment and compensation methods for unit nonresponse including the formation of weighting classes based on response propensities in different groups, as well as calibration methods, such as poststratification. In 1990, an international workshop on household survey nonresponse was initiated by Robert Groves, and this workshop still convenes annually; materials from the workshop are found on its website www.nonresponse.org.

Despite the development of methods for dealing with nonresponse, nonresponse rates increased considerably over the years in most countries. For instance, in 1970, the nonresponse rate in the Swedish Labor Force Survey was 2%, and currently in 2016 it is approximately 40%. However, a high nonresponse rate in isolation is not a solid indication of high nonresponse bias, since bias is also a function of the differences between respondents and nonrespondents with regard to the variables under study. As such, it is understood that sometimes nonresponse rates matter and sometimes not (Groves and Peychateva, 2008). Over the years, considerable energy has been devoted to developing methods that can help control the nonresponse rates and compensate for any residual nonresponse. Regardless, it is unlikely that in the foreseeable future, there will be any major declines in nonresponse rates, particularly given the recent proliferation of inexpensive high-technology modes of data collection.
Two common methods of compensating for item nonresponse were developed—imputation and multiple imputation, both of which replace missing data with modeled or simulated data. For instance, simple forms of “hot deck imputation” were first introduced at the U.S. Census Bureau in 1947. The principles for early uses of various imputation procedures are described in Ogus et al. (1965), but these principles differ considerably from those used today. Initially, the justification for using imputation methods was to create rectangular data sets by filling in the holes generated by missing data, since it was considered very difficult to handle missing data computationally. Consequently, the Census Bureau instituted very strict rules regarding the level of permissible imputations, whereby at most 2% item imputation was allowed, but if there were high demands on timeliness, this limit could be stretched to 5%. This is, of course, a far cry from today’s use of imputation where allowable rates are much higher given the increased sophistication and resulting accuracy of present-day methods.

Yet another source of nonsampling error that was identified early on was that survey staff such as interviewers, enumerators, and coders could generate both systematic and variable errors. Mahalanobis (1946) invented the method of interpenetration for estimating interviewer effects by suggesting the assignment of a random subsample of all interviews to each interviewer rather than an assignment based on practical considerations (i.e., assigning the interviews for all selected individuals in a primary sampling unit). For field interviewing, interpenetration was, of course, more costly than assignments based on practicality, but studies showed that individual interviewer styles could introduce a substantial cluster effect that could not be ignored. Interpenetration methods demonstrated that respondents within an interviewer assignment tended to answer in ways that were intrinsic to that specific interviewer’s style. Examples of style variation might include systematic deviations from the actual question wording or systematically inappropriate feedback on the part of the interviewers. Such errors could result in a correlated variance that is typically integrated as part of the total response variance but is not reflected in general variance estimates. Other operations mentioned earlier, such as coding, can also generate similar unaccounted for correlated variance, although they typically tend not to be large.

The topic of correlated variance is treated at length in Hansen et al. (1961) (see Section 1.3). Kish (1962) proposed an ANOVA model to estimate interviewer variance, and Bailar and Dalenius (1969) proposed basic study schemes to estimate the correlated variance components, of which interviewer effects is one (often substantial) part. It has been acknowledged that if survey conditions do not lend themselves to the ability to control interviewer errors, the effects can be dramatic. For instance, the World Fertility Survey Program has included cases of estimates whose variances were underestimated to an order of magnitude of 10 times, leading to strikingly understated margins of error (O’Muircheartaigh and Marckward, 1980). Unaccounted for correlated variance, such as in the aforementioned example, is the reason that standardized procedures have been instituted. Standardized procedures strive to ensure that interviewers, coders, and other groups work in the same way, thereby minimizing “cluster effects.” Observing interviewers in the field and monitoring of telephone interviews are means to control deviations from the standardized protocol.

Despite the ability to standardize procedures to minimize interviewer effects, other measurement errors were also prevalent and remain a concern. These measurement errors include errors due to questionnaire wording and questionnaire topics, general cognitive phenomena associated with memory and mode of data collection, and errors in field manuals. In fact, phenomena such as telescoping, memory decay, social desirability bias, comprehension, and

---

1 It is acknowledged that listwise deletion methods (where an entire record is excluded from analysis if any single value is missing) are rather easy to implement but suffer from bias and variance issues.
respondent fatigue were acknowledged relatively early on and discussed in the survey literature (Belson, 1968; Neter and Waksberg, 1964; Sudman and Bradburn, 1974). Even though most data collection agencies were aware that both measurement errors and processing errors could affect the quality of survey estimates, a substantial breakthrough did not occur until the release of the Jabine et al. (1984) report on Cognitive Aspects of Survey Methodology (CASM). The report emphasized the importance of measurement errors and their contributions to TSE, and defined response process models that have illuminated how some types of errors occur and how they can be mitigated. A response process model lays out the various cognitive steps a respondent undergoes from the survey participation request through to the delivery of his or her response. By disentangling these steps, it is possible to identify where the biggest risks are and how they should be dealt with. Response process models exist both for establishment surveys (Biemer and Fecso, 1995; Edwards and Cantor, 1991; Willimack and Nichols, 2010) and for surveys of individuals (Tourangeau et al., 2000).

The discussions and developments on controlling errors have followed different lines of thought over the years. For a large agency, such as the U.S. Census Bureau, rigorous controls of specific error sources were strongly advocated in the past. At the same time, there was a realization that extensive controls were expensive and their use had to be balanced against other needs. To the U.S. Census Bureau and other large producers of statistics, this imbalance was most obvious in the editing operation, which itself is a QC operation. Large amounts of resources were allocated to editing, which remains the case even today (de Waal et al., 2011). The purpose of these rigorous controls was to reduce biases and correlated variances, so that the TSE would consist mainly of sampling variance and simple response variance, both of which could be calculated directly from the data. This general strategy of controlling errors reduced survey biases to some extent. For example, nonresponse adjustments that take into account various response classes led to decreased nonresponse bias. Adherence to appropriate questionnaire design principles led to decreased measurement biases. Standardized interviewing, monitoring, and national telephone interviewing led to decreased correlated interviewer variance. But there still remain many biases that are generally not taken into account in current-day survey implementation.

The strategy of focusing on specific error sources to minimize the impact on the TSE has some inherent issues associated with it. First, rigorous controls are expensive and time-consuming, and additional control processes make most sense when the underlying survey process is under reasonable control to begin with. Second, the practice of investigating one error source at a time can be suboptimal. Some errors are more serious than others and this relative importance varies across surveys and uses. Third, all errors cannot be simultaneously minimized, since they are interrelated. For instance, in an attempt at reducing the nonresponse rate, we might induce an increased measurement error. Recent work on TSE has concentrated more on the simultaneous treatment of two or more error sources. For instance, West and Olson (2010) discuss whether or not some of the interviewer variance should really be attributable to nonresponse error variance. Also, Eckman and Kreuter (Chapter 5 in this volume) discuss interrelations between undercoverage and nonresponse. Fourth, in addition to survey errors and costs, Weisberg (2005) points out that sometimes errors cannot be minimized because correct design decisions are unknowable. For instance, asking question A before question B may affect the answers to question B, and asking question B before question A may affect the responses to question A. Therefore, it may be impossible to remove question order effects regardless of resources spent.

Thus, the approach aiming at reducing specific error sources is very important, but the error structures are more complicated than previously believed. Therefore, the inherent issues mentioned need to be addressed in more detail.

The second strategy toward the treatment of specific error sources uses evaluation studies as a means of quantifying the various sources of errors. Typically, evaluation studies are conducted
after the survey or census has concluded, and are a means of estimating the size of the total error or the error of the outcome of a specific survey operation, such as coding. Most well-known evaluation studies have been conducted in connection with U.S. censuses. A census is an ideal vehicle for studying survey processes and survey errors. The main methodology used is a comparison of the outcome of the regular survey or census compared to the outcome of a sample using preferred (but financially, methodologically, or administratively resource intensive) procedures or gold standard methodologies. Assuming that the gold standard is correct, the difference between the two is an estimate of the TSE, even though the difference is likely to either understate or overstate the true TSE. ASPIRE is a recent innovation used to evaluate TSE. It is an approach based on a mix of quality management ideas, as well as quantitative and qualitative assessments of the magnitude of TSE. This approach is further discussed in Section 1.4.

The evaluation programs conducted as part of the U.S. population censuses in 1940, 1950, and 1960 revealed important error and process problems, which led to significant procedural changes in future censuses and surveys. For instance, findings regarding the adverse effects of correlated variance induced by census enumerators as well as large costs associated with the enumeration process led to the decision to use more self-administration by mail in the census (U.S. Bureau of the Census, 1965). Currently, there is considerably diminished engagement in large evaluation studies mostly because of the enormous financial investments needed, but also because they are typically implemented long after they can be really helpful in any improvement work. For instance, the results of the evaluation of the coding operation in the 1970 U.S. Census were released in 1974 (U.S. Bureau of the Census, 1974b). Postenumeration surveys are still conducted in the U.S.A. to estimate the coverage error, and most importantly, how many people were missed in the census, since this so-called undercount can have a great impact on the distribution of funds to different regions in the country. In this case, the gold standard is a partial re-enumeration on a sample basis, where the estimation procedure resembles capture-recapture sampling (Cantwell et al., 2009).

1.3 Survey Models and Total Survey Design

During the period 1950–1970, much development was devoted to survey models aimed at providing expressions of the TSE as a combination of mean-squared-error (MSE) components. The U.S. Census Bureau survey model is perhaps the best known of these. In that model, the MSE of an estimate \( x \), \( \text{MSE}(x) \), is decomposed into sampling variance, simple response variance, correlated response variance, an interaction term, and the squared bias. In some versions of the model, there is also a component reflecting the relevance, which is the difference between the survey’s operational goal and its ideal goal. For instance, there is an operational definition of being employed used by official statistics agencies, which differs from an ideal definition that is more relevant but unattainable. The purpose of the survey model is to articulate the relative contribution to TSE from different components and to be able to estimate TSE more easily using components that can be added together. The model is described in numerous papers including Eckler and Hurwitz (1958), Hansen et al. (1961, 1964). The main issue with this model is its incompleteness in the sense that it does not reflect all the main error sources, most conspicuously, nonresponse and noncoverage. The model focuses solely on measurement errors and sampling errors. This is an obvious deficiency, specifically discussed by Cochran (1968). However, the model offers the opportunity to estimate errors beyond those induced by sampling and simple response variance. Although the above papers offer suggestions on how to estimate these components, Bailar and Dalenius (1969) provide a more comprehensive list of basic study schemes that could be used to estimate all components of error. The schemes use replication,
interpenetration, or combinations thereof. Some of these schemes are, however, rather elaborate and unrealistic. One scheme prescribes repeated reinterviews, which would be very difficult to implement given typical survey resource constraints. The estimation from these models of design effects due to variances associated with interviewers, crew leaders, supervisors, and coders has been particularly useful and has led to radical changes in census data collection procedures, as well as standardization and automation of other survey processes. Interviewer variance studies are relevant to many surveys, and more sophisticated schemes for its estimation are presented in Biemer and Stokes (1985).

The literature on survey models extends beyond that which comes from the U.S. Census Bureau. For instance, Fellegi (1964) introduced covariance components that Hansen and colleagues had assumed to be zero, including correlation of response deviations obtained by different enumerators (e.g., arising from specific features of training procedures) and correlation of sampling and response deviations within enumerators (e.g., the tendency for the same enumerator to induce different responses from elderly respondents than from young respondents).

Following Kish (1962), Hartley and Rao (1978) used mixed linear models to estimate nonsampling variance, and Lessler and Kalsbeek (1992) expanded the U.S. Census Bureau survey model by including also a component reflecting nonresponse. Bailar and Biemer (1984) made a similar attempt earlier but did not suggest specific estimates due to complexities relating to interaction terms.

In principle, the survey models and information about specific error sources can be used as inputs to survey designs. In this case, the aim is to develop a design such that the MSE is minimized given a fixed budget and any other constraints, assuming that all major sources of error are taken into account. A good design elucidates information about the relative impact of different error sources on the estimates, as well as the costs associated with reducing these effects. However, designs may vary for different survey estimates, and therefore, the use of the MSE should be considered as a planning criterion only. As Dalenius (1967) points out, “there is as yet no universally accepted survey design formula that provides a solution to the design problem and no formula is in sight.” Such a formula would have to take into account activities such as pretesting, implementation of operations, controlling operations, and documenting results. A formula did not exist in 1967 and still does not exist today.

A design approach toward the partial treatment of TSE suggested by Dalenius (1967) and Hansen et al. (1967) contained a number of steps that included the following:

- Specifying the ideal survey goal, which would permit an assessment of the relevance component;
- Developing a small number of alternative designs based on a thorough analysis of the survey objectives and the general survey conditions;
- Evaluating design alternatives with a view to understanding their respective preliminary contributions to the key components of the MSE, as well as their associated costs;
- Choosing an alternative design or some modified version of a design—or deciding not to conduct the survey at all;
- Developing an “administrative design” including components such as feasibility testing, a process signal system (currently called “paradata”\(^2\)), a design document, and a backup plan.

\(^2\) The paradata of a survey are data about the process by which the survey data were collected. The term first appears in Couper’s presentation of his paper (Couper, 1998) although the term is not present in the actual paper. Examples of paradata topics include the times of day interviews were conducted, interview duration, how many times there were contacts with each interviewee or number of attempts to contact the interviewee, and the mode of communication (such as phone, the web, email, or face to face). This definition is easily extended to encompass other survey processes (Lyberg and Couper, 2005) such as editing and coding.
Another approach to the treatment of TSE was suggested by Leslie Kish during an interview (Frankel and King, 1996). Influenced by the strong Bayesian-focused cadre at the University of Michigan in the 1960s, Kish suggested that Bayesian models be used to quantify some of the error components. Kish drew on the contributions by researchers such as Ericson (1969) and Edwards et al. (1963) regarding the use of Bayesian methods in survey sampling and psychometrics. Kish suggested that judgment estimates of biases could be combined with sampling variances to achieve more realistic and less understated estimates of TSE. Kish did not rule out the possibilities of using nonprobability sampling and Bayesian modeling to shed light on certain survey phenomena. Dalenius was also open to what he called “neo-Bayesian” ideas in survey sampling, and one paper he wrote discussed the use of diffuse priors in sample surveys (Dalenius, 1974). He commissioned Lyberg to write a review paper on the use of neo-Bayesian ideas in surveys (Lyberg, 1973).

Although Dalenius (1967) held a concept of total survey design that encompassed all known error sources, and Kish (1965) contemplated Bayesian ideas as input to survey design, these ideas did not materialize into a methodology that could be fully used at the time. This is because the treatment of all sources of error held too many unknowns and because Bayesian modeling was considered very demanding from a computational point of view at the time. Therefore, the TSE perspective lost some of its attraction during a relatively long period (between 1975 and 2000), because the survey model approach proved to be complicated, its components were computationally intractable, and the models were incomplete. No agency really attempted to estimate TSE, with the exception of Mulry and Spencer (1993), who tried to estimate the total MSE of the 1990 U.S. Census. Instead survey organizations continued to work on methods that could reduce specific error sources as a consequence of a rapid development of new modes, combinations of modes, and methods for handling cognitive aspects of surveys. Near the end of the era of “disinterest,” Forsman (1987) expressed disappointment with the small role that survey models had played in survey implementation to date. At roughly the same time, Biemer and Forsman (1992) showed that basic reinterview schemes did not work as intended, and Dillman (1996) was concerned about the lack of innovation within the U.S. Federal Statistical System with respect to addressing these issues. Finally, Platek and Särndal (2001) posed the following question: “Can a statistician deliver?” voicing their concern regarding the theoretical foundations of survey methodology, which included the topic of TSE. The Platek and Särndal article came to serve as a wake-up call for parts of the survey industry. A new workshop, the International Total Survey Error Workshop (ITSEW), convened its first meeting in 2005 and has, since 2008, met annually. The purpose of the workshop is to promote TSE thinking as well as to encourage studies that aim at joint investigations of more than one error source.

1.4 The Advent of More Systematic Approaches Toward Survey Quality

Around 1970, there was general agreement among prominent survey organizations that all main error sources ought to be taken into account when designing surveys. A few years earlier Hansen, Cochran, Hurwitz, and Dalenius had decided to write a textbook on total survey design, but the plan was abandoned due to the sudden demise of Hurwitz in 1968 (T. Dalenius, Personal communication with Lars Lyberg, 1968). Eventually, Groves (1989) wrote a seminal textbook along these lines.

One of the problems with the work on survey errors during that era was the absence of a process perspective and a consideration of continuous improvement. For instance, improvement
work was concentrated on measuring and decreasing errors, often without considering a process perspective. The user of statistics was a rather obscure player and even though there were user conferences (Dalenius, 1968), information about errors and problems flowed in one direction, namely from producer to user. Users were rarely asked to provide feedback to producers in this regard. Statisticians sometimes “role-played” as subject-matter specialists during the design phase of surveys, rather than engaging such specialists directly. Even though industrial process control had been used extensively at the U.S. Census Bureau and other places, no real process thinking was embedded in the strategies to reduce errors. Some consideration was given to process signal systems functioning as early warning systems, much in the same vein as paradata do today. However, continuous improvement of survey processes was not well developed, and when problems occurred, “rework” was a common remedy.

In roughly 1980, quality management and quality thinking become popular in organizations. Quality management developed as a science (Drucker, 1985) and quality systems such as total quality management (TQM) and Six Sigma entered the scene. Statistical organizations jumped on the bandwagon for two reasons.

First, there was pressure to recognize the user in more formalized ways, because of the acknowledgment that for statistics to be relevant they had to be used (Dalenius, 1985). Previously, attempts such as the “U.S. Standard for Presentation of Errors” (U.S. Bureau of the Census, 1974a) and error declarations in connection with the release of survey reports were quite technical and were developed without much contact with users. The era had arrived when the user was recognized as the customer or the representative of a paying customer, both who had the right to achieve value for money. The second reason for introducing quality management principles was cost. The production of statistics was expensive, and without process changes that resulted in more cost-effective outputs, competitors might take over.

There are several activities related to the principles of quality management, which became important in the production of statistics. Flow-charting of processes, plotting of paradata on control charts, and using cause-and-effect diagrams are examples of activities that became popular within the process improvement paradigm. There was an acknowledgment of a complementarity between survey quality and survey errors. It was recognized that accuracy could not be considered the sole indicator of survey quality, in the same way that the nonresponse rate cannot be considered the only indicator of accuracy of a survey. Dimensions other than relevance and accuracy were identified as important to users, most notably the dimensions of accessibility and timeliness, in an acknowledgment that accurate statistics might have limited utility if difficult to access or received too late. Considerable development was invested in a number of quality frameworks that articulated the various dimensions of quality. The first framework was produced by Statistics Sweden (Felme et al., 1976), and since then a number have followed. For instance, the Organisation for Economic Co-operation and Development’s (OECD) 2011 framework has eight dimensions: relevance, accuracy, timeliness, credibility, accessibility, interpretability, coherence, and cost-efficiency. Eurostat, the statistical agency of the European Statistical System, has developed a Code of Practice that contains 15 different dimensions that relate to quality (Eurostat, 2011).

Statistical organizations have changed as a result of the global quality movement. Many organizations now use customer satisfaction surveys, process control via paradata (Kreuter, 2010), organizational quality assessment using excellence models such as Six Sigma (Breyfogle, 2003), quality improvement projects (Box et al., 2006), and current best methods (Morganstein and Marker, 1997). In 2008, Statistics New Zealand submitted a proposal for a new Generic Statistical Business Process Model (GSBPM) (Statistics New Zealand, 2008), which defines phases and subprocesses of the statistical lifecycle. The model has gradually been refined and its fifth version was released in 2013 (see Figure 1.1). The GSBPM is intended to apply to all activities...
<table>
<thead>
<tr>
<th>Quality management/metadata management</th>
<th>Specify needs</th>
<th>Design</th>
<th>Build</th>
<th>Collect</th>
<th>Process</th>
<th>Analyze</th>
<th>Disseminate</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Identify needs</td>
<td>2.1 Design outputs</td>
<td>3.1 Build collection instrument</td>
<td>4.1 Create frame &amp; select sample</td>
<td>5.1 Integrate data</td>
<td>6.1 Prepare draft outputs</td>
<td>7.1 Update output systems</td>
<td>8.1 Gather evaluation inputs</td>
<td></td>
</tr>
<tr>
<td>1.2 Consult &amp; confirm needs</td>
<td>2.2 Design variable descriptions</td>
<td>3.2 Build or enhance process components</td>
<td>4.2 Set-up collection</td>
<td>5.2 Classify &amp; code</td>
<td>6.2 Validate outputs</td>
<td>7.2 Produce dissemination products</td>
<td>8.2 Conduct evaluation</td>
<td></td>
</tr>
<tr>
<td>1.3 Establish output objectives</td>
<td>2.3 Design collections</td>
<td>3.3 Build or enhance dissemination components</td>
<td>4.3 Run collection</td>
<td>5.3 Review &amp; validate</td>
<td>6.3 Interpret &amp; explain outputs</td>
<td>7.3 Manage release of dissemination products</td>
<td>8.3 Agree on action plan</td>
<td></td>
</tr>
<tr>
<td>1.4 Identify concepts</td>
<td>2.4 Design frame &amp; sample</td>
<td>3.4 Configure workflows</td>
<td>4.4 Finalize collection</td>
<td>5.4 Edit &amp; impute</td>
<td>6.4 Apply disclosure control</td>
<td>7.4 Promote dissemination products</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5 Check data availability</td>
<td>2.5 Design processing &amp; analysis</td>
<td>3.5 Test production system</td>
<td></td>
<td>5.5 Derive new variables &amp; units</td>
<td>6.5 Finalize outputs</td>
<td>7.5 Manage user support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6 Prepare business case</td>
<td>2.6 Design production systems &amp; workflow</td>
<td>3.6 Test statistical business process</td>
<td></td>
<td>5.6 Calculate weights</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.7 Finalize production system</td>
<td></td>
<td>5.7 Calculate aggregates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.8 Finalize data files</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.1** The generic statistical business process model. Source: Statistics New Zealand (2008).
undertaken by producers of official statistics. It can be used to describe and assess process quality independent of data sources used. A more complete description of the impact of quality management principles on survey organizations is given in Lyberg (2012).

Biemer (2010) formally defined the TSE paradigm as part of a larger design strategy that sought to optimize total survey quality (TSQ) and that included dimensions of quality beyond accuracy. The dimensions under consideration could be user-driven, and could be adopted from an official framework of the kind mentioned above or from any quality vector specified by the user. The basic elements of the TSQ paradigm include: design, implementation, evaluation, and the assessment of the effects of errors on the analysis. In the design phase, information on TSE is compiled, perhaps through quality profiles, which are documents containing all information that is known on the survey quality. From this, the major contributors to TSE are identified and resources are allocated to control these errors. During the implementation phase, processes for modifying the design are entertained as a means of achieving optimality. The evaluation part of the process allows for the routine embedding of experiments in ongoing surveys to obtain data that can inform future survey designs.

In relation to the first two pillars of the paradigm (design and implementation), a number of strategies have been developed that allow for design modification or adaptation during implementation to control costs and quality simultaneously. The activities in support of these strategies are conducted in real time and the strategies include continuous quality improvement, responsive design, Six Sigma, and adaptive total design.

The first strategy, “continuous quality improvement,” is based on the continuous analysis (throughout implementation) of process variables, process metrics, or paradata that have been chosen because stable values of them are critical to quality. As a result of the analysis, specific interventions might be deemed necessary to ensure acceptable cost and quality.

A second strategy, called “responsive design” (Groves and Heeringa, 2006), was developed to reduce nonresponse bias. It is similar to continuous quality improvement but includes three phases: experimentation, data collection, and special methods to reduce nonresponse bias.

A third strategy is the use of the Six Sigma excellence model. It emphasizes decision making based on data analysis using a rich set of statistical methods and tools to control and improve processes. Six Sigma is an extreme version of continuous improvement.

A fourth and final strategy, called “adaptive total design and implementation,” is a monitoring process which is adaptive in the sense that it combines features of the previous three strategies. In all these strategies, the analysis of metrics is crucial. The theory and methods for industrial QC can be used (Montgomery, 2005) in the same way as they were during the U.S. Census Bureau operations in the 1960s. However, what differs is the treatment of different kinds of variations. Process variation used to be attributed solely to operators, for instance, while the current prevailing philosophy is that it is the underlying processes themselves that more often have to change.

The third pillar of the paradigm is the TSE evaluation. Such an evaluation can address any dimension of survey quality and is essential to long-term quality improvement. Examples include nonresponse bias studies and measurement bias studies. Of particular importance is the consideration of the joint effects of error sources and their interactions, rather than just single sources of error such as nonresponse.

The fourth pillar is the assessment of the effects of errors on the analysis. This is a neglected area but has been discussed in the literature by Biemer and Stokes (1991), Koch (1969), and Biemer and Trewin (1997). (See also Chapter 23 in this volume.) The effects of errors depend on the kind of parameter that is estimated and also on the specific use of the deliverables.

It was mentioned earlier that both users and producers of statistics alike have problems understanding the complexity of TSE and its components. Some types of errors are difficult to explain,
and therefore there is a tendency to emphasize errors and concepts that are easily understood, such as nonresponse. Furthermore, this lack of understanding is exacerbated by the fact that statistical agencies do not attempt to estimate TSE at all. However, recently the ASPIRE system (A System for Product Improvement, Review, and Evaluation) was developed at Statistics Sweden by Paul Biemer and Dennis Trewin in an attempt to assist management and data users in assessing quality in a way that can be easily understood. In this system, the MSE is decomposed into error sources. A number of somewhat subjective criteria on (among other things) risk awareness, compliance with best practice, and improvement plans are defined and quality rating guidelines are defined for each criterion. Rating and scoring rules are defined, and risk assessments as well as an evaluation process are performed. ASPIRE is described in Biemer et al. (2014) and has been successfully used for the 10 most critical products at Statistics Sweden; the quality of these products has improved over the four rounds conducted thus far.

Moving beyond the concept of TSQ, the concept of total research quality (TRQ) was introduced recently by Kennet and Shmueli (2014). The authors penned the term “InfoQ” to describe attempts at assessing the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or data mining.

1.5 What the Future Will Bring

The survey landscape is currently transforming quickly. Because traditional surveys are costly and time-consuming, they are being replaced or complemented by other types of information sources. “Opt-in online panels” based on nonprobability sampling methods borrowed from the presampling era are used to create representative miniature populations and have become quite common, especially in marketing and polling firms. The panels consist of individuals who have been recruited by banners on a website, or by email—and who have provided their email addresses to the implementing firm. Double opt-in online panels means that the recruited individuals receive a response from the firm and are asked to confirm their willingness to participate as well as to provide their email address and other personal information. Sometimes those who join receive an incentive. There is even an ISO (2009) standard for using and maintaining such panels, sometimes called “access panels,” but as of the present, there is no theory to back the use of such panels. However, it is not uncommon to find that the results based on these panels produce outcomes that are quite similar to those using probability sampling (AAPOR, 2010; Wang et al., 2015), although it is often impossible to disentangle the magnitude of the differences. Online panels based on opt-in and double opt-in are likely here to stay, but data quality issues in relation to these have yet to be resolved. The use of Bayesian modeling (Gelman et al., 2014) is a possible route to explore, as well as the sensible adjustments of nonprobability samples using multilevel regression and poststratification, as demonstrated by Wang et al. (2015) in election predictions.

Some research fields use survey procedures without adopting a TSE perspective. Big Data allow for the harvesting and analysis of sensor data, transaction data, and data from social media. As shown in the recent AAPOR (2015) task force on Big Data and in Chapter 3 in this volume, it is possible to develop a TSE framework for Big Data. Hard-to-sample populations and international comparative surveys are other examples of survey areas that have their own research traditions (Chapter 9 in this volume; Tourangeau, 2014) that could benefit from a TSE perspective, and such work is underway. The use of administrative data also needs its own TSE framework (Wallgren and Wallgren, 2007). Even data disclosure limitation can be viewed from a TSE perspective (Chapter 4 in this volume).
Figure 1.2  Subjective sample of events in the evolution of the concept of TSE.
It is heartening to see that quality issues have resurfaced as an area of interest for survey methodologists and data users alike. Recently, media outlets, who are important users of data, have developed publication guidelines including criteria on response rate, question wording, sampling method, and sponsorship. The New York Times, The Washington Post, and Radio Sweden are examples of such outlets. This is part of a greater trend toward data-driven journalism that is based on analyzing and filtering large data sets for the purpose of creating news stories based on high-quality data.

A new survey world that uses multiple data sources, multiple modes, and multiple frames is at our disposal, and it is essential that quality considerations keep pace with such developments to the extent possible. Indeed, promoting and defending ideas on data quality and sources of error is an important, albeit daunting task.

In closing, Figure 1.2 provides the authors’ subjective summary timeline of some of the most important developments in TSE research from 1902 to present day.

References


References


