Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 25–32

Comment

Paul P. Biemer[1]

Introduction

To sum up Platek and Särndal's response to the question posed in the title ''Can a statistician deliver?,'' the authors seem to be implying ''Yes he or she probably can, but have not delivered much in the past 40 years.'' Although some of us might bristle at such pointed criticism, to some extent we must admit they are correct: statisticians have made slow progress toward developing the total survey error modeling concept envisioned by Morris Hansen and his colleagues. Look at any statistical report published by a government agency and you are much more likely to find only standard errors of the estimates rather than total mean squared errors. It does appear that not much progress has been made since the late 1950's and early 1960's on informing data users of the real levels of uncertainty in estimates arising both from sampling and from nonsampling sources.

Some might argue that such criticism is unfair since it is predicated on an impractical standard and the authors are measuring progress against an unattainable or unrealistic ideal. Is it reasonable to ever expect that systematic and variable errors associated with nonresponse, measurement error, and coverage error will be routinely assessed and reported in survey work? If not, what, then, should be the vision for assessing, controlling, and reporting nonsampling errors in surveys and how can we as a profession begin to achieve this vision? In this comment, I will share my thoughts in response to these questions and consider the progress that statisticians have made regarding the total error modeling concept.

Truths and Myths in Total Survey Error Evaluation

In defining a new vision for the future regarding total survey error, a reasonable first step is to consider what data on survey quality are needed and for what purposes. Once this is understood, we can see where the field stands and where it still needs to go.

The nonsampling error literature suggests a number of uses for estimates of total survey error components and the authors reiterate most of these. These uses are often cited to support the need for more quality evaluation studies and to justify the expense of such studies. In reality, however, the statisticians who conduct the evaluation studies and report the results are usually not the same statisticians who put these evaluation results to actual use. This disconnect may have led to some unintentional ''over-selling'' of the total survey error products in some cases and to not providing the needed information on survey error in others. Below we consider some of the uses cited for nonsampling error evaluations and some of the myths surrounding these uses.

[1] R T I, Research Triangle Park, NC, U.S.A.

© Statistics Sweden

1. Comparing the accuracy of data from alternative modes of data collection or estimation methods.

Perhaps the most important use of total survey error estimation is to assess the relative quality of alternative data collection modes (for e.g., mail, telephone, face to face, etc.) or for comparing alternative estimation methods. For example, a survey methodologist may wish to compare the accuracy of health data collected by mail and by telephone. Typically, a mode comparison study would be conducted that is based upon a split ballot design where some portion of the sample is collected by telephone and the remaining portion is collected by mail. While this may be sufficient for deciding whether the two modes give different or the same results for some characteristics, it is usually not sufficient for determining which mode is better for the key items in the survey. For this purpose, the total MSE's of the estimates from both modes must be compared.

However, even this may not provide adequate information on the data quality from each mode if it is important to determine whether the differences in the estimates are due primarily to nonresponse bias, measurement bias, coverage bias, or some other bias. For example, response rates may differ according to the organization conducting a survey. This is part of the so-called house effect. Thus, if the primary cause of the mode differences is nonresponse, then the mode effects may vary by the organization collecting the data. In that case, a decomposition of the total MSE is required in which the major components of the MSE are estimated individually for each mode. Unfortunately, this is seldom done in mode comparison studies, a fact which partially explains why the survey methods literature is plagued with inconsistent results across studies.

2. Optimizing the allocation of resources for the survey design.

Design optimization is another reason often cited for estimating the magnitudes of the nonsampling error components; in fact, Platek and Särndal mention this as one of the primary uses of nonsampling error estimates. For this purpose, the survey designer would like to know how much of the total survey error is contributed by each of the major sources of error in a survey, such as: nonresponse, frame coverage, the questionnaire, the interviewer, the mode of interview, the respondent, data editing, and so on. In practice, however, this information by itself is seldom sufficient since in order to choose between alternative survey designs and implementation methods, the designer needs to know not only the magnitude of the error contributed by a particular source, but also how this error is affected by the various design choices that are feasible for the survey.

For example, a survey designer may ask, ''Will nonresponse bias be reduced more by increasing the interviewer training budget by 10,000 USD or by allocating these funds to the nonresponse follow-up operation or by using them to increase the amount of incentives paid to respondents?'' Another designer may ask, ''Is it better to administer the survey using face to face or telephone interviewing, even though a 40 percent reduction in sample size is required to afford face to face interviewing.''

Although information on the components of nonsampling error may be known for a survey, this information by itself may have little utility for the survey designer. Determining the best allocation of survey resources to reduce nonsampling error requires information not only on nonsampling error components but also on how these are jointly affected by

the many allocation alternatives that the designer may consider. Such information usually does not exist for a survey. Indeed, it may be unreasonable or impractical to expect that it ever could be made available for a single survey with any regularity. Fortunately, it may not be necessary to compile such a vast database of cost, error, and method information for every survey to be optimized.

As an example, optimal strategies for mail survey design have been developed by using the results of experiments across many surveys on a wide range of topics. Using meta-analysis and other techniques for integrating this vast collection of research results, survey methodologists have identified what appears to be the ''best'' combination of questionnaire design and implementation techniques for maximizing response rates, minimizing measurement error, and reducing survey costs. This ''tailored design method'' for mail surveys is a good example of using a total survey error model to develop a theory and practice for ''optimal'' mail survey design. However, this approach is very different than what was envisioned by the early developers of the total error concept.

A similar comprehensive design approach has not yet been developed for the interviewer assisted modes, although there exists a vast literature covering most aspects of these designs. For example, there is literature on the relationship between length of training, training costs, and interviewer variance, but it is not known whether these relationships are transferable from one survey to another. There is also a considerable literature relating nonresponse reduction methods such as follow-up calls and incentives to response rates, and in some cases to nonresponse bias. Perhaps the total survey error concept that led to a theory of optimal mail survey design may one day be employed in the development of a theory and methodology for optimal face to face or telephone survey design.

3. Reducing the nonsampling error contributed by specific survey processes.

Information on the magnitudes of nonsampling error components contributed by specific survey operations can also be used to identify faulty operations that are in need of improvement.

However, rarely is this information sufficient for determining the causes of the errors – an essential step for effective error reduction. As an example, we may determine that interviewer variance is an important component of the total error for some key characteristics in a survey; however, this information alone is usually not adequate for determining the causes of the problem, be it interviewer training, the interview guidelines, the design of the questionnaire, an interaction between interviewer and respondent characteristics, or some other causes. Identifying the root causes of interviewer variance would require additional experimentation and testing.

The lack of information on the magnitudes of the error from specific sources has not necessarily stymied progress on improving most survey operations. For example, the application of cognitive methods for evaluating survey questions was a major innovation that has led to the reduction of many types of measurement errors arising from the questionnaire. Other notable innovations include the use of centralized telephone interviewing (reduction of interviewer variance), computer assisted interviewing (CAI) methods (reduction of the number of missing items and inconsistent reporting), audio computer assisted interviewing (ACASI, reduction of the deliberate misreporting of sensitive topics), and the use of prepaid incentives (reduction of nonresponse).

Platek and Särndal lament the lack of a unified theory for surveys and offer the field of physics as an exemplar for such a theory. However, surveys are as complex as the people and other entities that are surveyed. A unified theory for survey design seems as improbable as a unified theory for predicting human behavior across a myriad of situations. What appears to be emerging in the field, instead, is not a single theory but a collection of many related theories which deal with all aspects of survey design and which are somewhat specific to the survey topic and the population to be surveyed. Some examples of such theories are:

- factors that influence the decision by respondents to participate in a household survey,
- use of incentives as gifts rather than remuneration,
- use of graphics and imagery to aid respondents in navigating through a paper and pencil questionnaire,
- the order and context of questions as they affect measurement error, and
- best approaches for training and instructing interviewers for the reduction of interviewer variance.

4. Providing information to data users regarding the quality of the data or the reported estimates.

Platek and Särndal state that data on total survey error is needed ''in order to provide the user with objective information on the relative importance of different errors'' and that this information will aid user's understanding of the limitations of the data. This is true to some extent, but not as much as we would hope. For example, measures of nonsampling error indicating excellent or very good data quality create high user confidence in the quality of the data, while measures that imply only fair to poor data quality tend to have the opposite affect. In the end, many users still remain confused as to exactly what the measures of total error indicate about their specific uses of the data or how they should interpret the results of their analysis.

As an example, a report on survey quality may contain estimates of nonresponse bias for the key estimates – usually, means, totals, and proportions – produced from the survey data. This information is quite informative for assessing the accuracy of the prevalence estimates and the estimates of totals that may be of interest to the user. However, information on these estimates is inadequate if the user wishes to know how the nonsampling biases affect a logistic regression analysis or even a simple comparison of two estimates from the same survey.

Likewise, estimates of test–retest reliability may be provided in the data quality report which are useful for understanding the amount of variable error or response inconsistency in the data. Still the user is left to determine how the reported levels of reliability affect the results of a categorical data analysis or some percentile estimate. Beyond its ability to either instill trust or create distrust in the data, simply reporting nonsampling error components and MSE's in a data quality report is usually not sufficient for most forms of secondary data analysis. Yet many statisticians believe that this type of reporting is the ultimate solution for informing users about the limitations of the data.

Many statisticians would probably agree with Platek and Särndal's statement that ''to measure total survey error in an estimate is surely what statisticians ought to do.''

However, in actual practice the strongest case that can be made for estimating total survey error components is that it is useful for deciding between alternative modes of data collection or estimation methods (Purpose 1 above). For the other purposes, such information is either inadequate (Purposes 2 and 4) or not necessary (Purpose 3).

This is not to suggest that work on total error modeling estimation should be curtailed in any way, but perhaps to suggest that the ''bar'' be raised for the total error modeling concept. Measurement and reporting of nonsampling error is not an end unto itself and it may not even be the only means toward an end. Rather, understanding the causes and the prevention of nonsampling error is the key and should receive the highest priority. For some error components, this is more likely to involve interviewing a small representative sample of the target population using cognitive interview methods than a large study aimed at estimating a bias component. However, small-scale laboratory investigations used in conjunction with large-scale error component evaluation studies may be ideal for most purposes. Evaluation studies aimed at describing the effect of alternate design choices on total survey error are also extremely important since without them total survey design optimization is not possible.

Impediments to Realizing the Total Survey Error Concept

The original developers of the total survey error concept envisioned a time when all major components of sampling and nonsampling error would be considered in the design of surveys and routinely incorporated in the reported estimates of uncertainty. This has not happened, as Platek, Särndal, and other authors note. As they further note, we instead find that there is:

A. No routine measurement of the major MSE components other than sampling error.
B. Too little research on integrated modeling and joint estimation of survey error.
C. Not enough attention paid to nonsampling error by sampling statisticians.
D. No standardization of survey methods across statistical agencies.

Notwithstanding this lack of progress in the area of nonsampling error evaluation, a number of relatively recent innovations have resulted in significant data quality improvements. As previously noted, cognitive methods and meta-analyses of results in the survey methods literature have advanced mail and telephone surveys and have provided a much better understanding of error for these modes of data collection. In addition, new theories have been developed for guiding the design of mail surveys; understanding the complex relationships between nonresponse and the environment, the survey design, the interviewer and the respondent; and understanding how question position, context, and wording affect measurement error.

Progress to improve survey quality has also come through technological innovations such as computer assisted survey information collection (CASIC), an area which includes the use of many computer devices, automated routines, and computer assisted methods for collecting, capturing and editing survey data.

There have been important innovations in the area of statistical error modeling as well. The authors note the advances in post-survey adjustment methods for compensating for nonresponse and noncoverage, but fail to mention the importance of latent class analysis

and multilevel modeling to the field. Statisticians are also beginning to recognize the importance of other psychometric methods to the study of nonsampling errors, such as correspondence analysis and item response theory, including Rasch models. Quality profiles have been developed for a number of important surveys in the U.S. and the trend toward periodically producing this type of nonsampling error summary for critical national surveys is increasing. Admittedly, our progress still falls short of the vision implied by the early developers of the total error modeling concept, but as mentioned at the start of this comment, there may be some good reasons for that.

Regarding (A), one can cite the many problems inherent in measuring nonsampling error components as a serious impediment to progress in this area. Estimating bias requires a set of gold standard measurements (i.e., the truth) and true values are often impossible to measure accurately. Measuring interviewer variance using interpenetrated assignments is quite difficult to do in a face to face survey and the estimates are often unstable unless many interviewers are involved. Even the simple test–retest reinterview study can be quite problematic methodologically. Reinterview nonresponse, response conditioning (or carry-over) effects, and the lack of independence between interview and reinterview response errors tend to erode our confidence in reliability estimates.

In addition, many survey sponsors do not perceive many benefits investing scarce survey resources in data quality evaluations or else feel that the benefits do not compensate for the costs and risks of diverting resources away from the primary data collection activities. Too often evaluation studies have stopped with the measurement of one or more MSE components rather than continuing until a thorough understanding of the causes of the errors has been acquired. Consequently, attempts at revising the survey designs on the basis of the evaluation results have not been effective at reducing survey errors.

Further, there does not seem to be much demand from the user community for information on data quality. Perhaps the majority of users find little use in estimates of MSE components for most types of analysis they perform. As noted earlier, quality profile reports usually provide estimates of the bias in point estimates which may be of little use when the analysis goes beyond point estimation. In addition, many users lack the statistical training needed to interpret the results of data quality assessments and, thus, tend to ignore the evaluation findings even when they are provided.

The problems noted in (B) and (C) above could arise largely as a result of lack of courses that deal with survey nonsampling error in university statistics departments. In the U.S., such courses can be found in only a handful of universities that offer degrees in survey methodology with heavy focus on statistics. Otherwise, courses in nonsampling modeling and estimation are extremely rare worldwide. Thus, it is not surprising that few dissertations are written in this area and that few statisticians choose survey error modeling as their primary research area when they join the research community.

Finally, the problem noted in (D) above would seem to be a product of the lack of progress already noted in (A)–(C). However, most survey methodologists would probably agree that methods for conducting mail surveys are fairly standard across many statistical agencies in the U.S. This progress is owed to the efforts of survey methodologists such as Dillman (1978; 2000) who have integrated the findings from a diverse body of literature on mail survey techniques to arrive at a standardized approach for conducting surveys by

mail. A universal standard for telephone, face to face, or Web-based surveys, on the other hand, does not exist.

The Need for a Revised Total Error Concept for Surveys

The somewhat condemnatory quote in Smith (1990) and reiterated by Platek and Särndal correctly notes the lack of progress by statisticians toward the total survey error modeling concept even after 50 years of intensive research. Perhaps this suggests that the 1950's concept itself should be reexamined. For all the reasons described above, the routine reporting of nonsampling error components in surveys does not seem to be plausible or even desirable for most purposes. A revised concept or vision is needed for the 21st century.

It is interesting to consider how this revised concept might differ from the original one. Certainly it should take into account the evaluation and design optimization tools that were not available until recently. In addition, the new concept might be tailored to the specific need it is intended to fulfill since such needs may vary by survey. The following is a list of features that might be desirable in this new concept for total survey error.

For comparing the accuracy of alternative data collection modes or estimation methods,

- Compare not only the total MSE for each alternative, but also each of the major components of error, including: nonresponse bias, coverage bias, measurement (or mode) bias, simple response variance, and interviewer variance.

For optimizing the allocation of survey resources,

- Provide data on the magnitudes of the nonsampling biases and variances from the major sources of error or survey operations as well as information on how the magnitudes of these errors are affected by design alternatives having the greatest impact on survey costs.
- Rather than relying solely on the results that are specific to the particular survey to be optimized, employ meta-analysis and other study-integration approaches.
- Using empirical results as a guide, develop theories for the optimal design of specific survey operations.
- Document these approaches so that they can be tested and critically reviewed by methodologists across statistical agencies and, if warranted, adopted as standard practice.

For reducing errors contributed by specific survey processes,

- Estimate the major components of nonsampling error periodically for critical national surveys in order to identify the major sources of nonsampling error or rely on the results from the survey methods literature to suggest survey processes that contribute the least to total survey error.
- Use cognitive methods, small-scale laboratory or field experiments, respondent and interviewer debriefing methods, etc. to determine the root causes of the errors and identify innovative methods for reducing the errors.

- Implement the revised methods and evaluate the change in data quality for the operation.
- Document the results for internal and external standardization purposes.

For providing data users with information on data quality.

- Publish estimates of the MSE components for the survey as well as information to help users understand how the nonsampling errors in the data affect some of the major uses of the data. This could include illustrations showing the effects of bias and variance components on point estimate comparisons, correlation coefficients, contingency table analysis, regression analysis, and quantile estimates.

For this revised total error concept, I think the answer to the question posed in the title ''Can a statistician deliver?'' might be, ''Yes, but there is still much work to be done'' – a much more favorable response to a much more reasonable question.

Congratulations to the authors for producing this very thought-provoking article and my thanks to the editor of JOS for allowing me this opportunity to comment on it.

References

Dillman, D. (1978). Mail and Telephone Surveys. Wiley.

Dillman, D. (2000). Mail and Internet Surveys. Wiley.

Smith, T. M. F. (1990). Comment on Rao and Bellhouse: Foundations of Survey Based Estimation and Analysis. Survey Methodology, 16, 26–29.

Received December 2000