International Journal of Public Opinion Research Vol. 23 No. 4 2011

doi:10.1093/ijpor/edq052 Advance Access publication 19 July 2011

Refining the Total Survey Error Perspective

Tom W. Smith

NORC, University of Chicago, IL, USA

Abstract

Total survey error (TSE) is a very valuable paradigm for describing and improving surveys, but it can be improved. First, either TSE needs to be limited to covering just instances of differences between true and measured values or TSE should be rechristened as total survey measurement variation (TSMV) if other forms of measurement-related variation are to be included. Second, the TSE/TSMV typology needs to be as detailed and comprehensive as possible. Third, TSE needs to be thought of as heavily involving the interaction of error components and the concept of comparison error should be used to extend TSE to cover multiple survey types. Fourth, the minimizing of TSE is an important goal in survey research and the TSE paradigm can be used as both an applied application and a research agenda to achieve that goal. Finally, TSE has both individual and aggregate components and an absolute and situational aspect. The role of each of these needs to be kept in mind.

Total Survey Error versus Total Survey Measurement Variation

Total survey error (TSE) is the sum of all the myriad ways in which survey measurement can go wrong (Smith, 2005). As Judith Lessler (1984, p. 405) notes, it is "the difference between its actual (true) value for the full target population and the value estimated from the survey [...]." Under this definition, TSE only refers to differences between true values and measured values. But as commonly applied, the TSE paradigm is used to cover not only differences between the true and measured values, but also differences in true values or for comparing different true values. For example, Groves (1987, p. S165) has noted in regard to "measurement error arising from the

All correspondence concerning this article should be addressed to Tom W. Smith, NORC, University of Chicago, 1155 East 60th Street, Chicago, IL, 60637, USA. E-mail: smitht@norc.uchicago.edu

questionnaire'' that ''most current research is examining the effects of question order, structure, and wording and does not purport to investigate the measurement of error properties of questions. Instead, researchers note changes in response distributions associated with the alterations.''

An analogy to the physical sciences can help to clarify the distinction. The boiling point of water (BPW) is conventionally defined as 212° Fahrenheit. But that is based on certain conditions applying. In particular, it assumes that the water is at sea level. In Chicago, the BPW is 211° and in Denver it is 202°. That is, the BPW lowers as altitude increases due to lower atmospheric pressure. There is no error in measuring the BPW across these localities, but variability in the true value due to differences in elevation and atmospheric pressure.

The broader concept might be called total survey measurement variation (TSMV) which includes both TSE and true variation due to differences in measurement. One needs either to use TSE in a more restricted manner consistent with the true versus measured-values definition or to expand TSE into TSMV by including measurement-related true variation.

Of course, the difference between error and nonerror variability is not always clear. In the case of question wordings, a well-defined concept of car ownership might be measured more reliably and accurately by various single and/or multiple questions than by others. These questions would differ in their ability to record the correct, car ownership status for a person (Smith, 1989). Alternatively, a series of related questions may tap the same general concept without measuring the same true value. For example, a series of questions about U.S. entry into World War II (Cantril, 1940, 1947) showed considerable variation in support for the United States becoming involved. Rather than seeing the differences as representing error in measuring a single true value, it makes more sense to acknowledge that the issue was complex and that the different questions addressed related, but not identical, issues and that the differences represented variability due to the different focus of the questions rather than error in measuring a single true value.

Another example is the large difference that occurs when asked about support for government spending for ''welfare'' versus ''assistance to the poor.'' On the 2008 General Social Survey (GSS) 25.4% wanted more spending for ''welfare,'' while 70.2% backed higher spending for ''assistance to the poor'' (Smith, 2009b). This large difference has consistently appeared for over 20 years (Smith, 1987, 2006b). If one sees the two questions as measuring the same true value, then one or both are in error and the example clearly fits under the TSE model. But if one sees these as measuring related, but not identical values, then their variation would not be covered by TSE, but would be covered under the TSMV paradigm.

As another example, context or order effects would generally be seen as representing variability rather than error (i.e., the different context produces

different true values on the subsequent questions). But if one context contributed to more mistakes or misunderstandings of the follow-up question, then the differences due to context might be seen as differences in true and measured values. (For a discussion of under what contexts measured values might be closer to measuring socially meaningful attitudes on subsequent questions, see Smith, 1991.)

The distinction between TSE involving differences between true and measured values and TSMV involving differences between different, but related, true values is somewhat complicated when dealing with attitudes and other introspective measures for which there is no external, objective way to ascertain the true value of a variable. In this case, the true value is undocumentable, but it can be thought of as the response that a respondent would give if there were no mismeasurement. There would be a true value even for questions on the most difficult, obscure, or unconsidered topics. Of course, the true value for such items might often be "Don't Know" and/or be very labile.

Classifying TSE

Bias and Variance

TSE comes in two varieties: (a) Variance or variable error which is random and has no expected impact on mean values, and (b) bias or systematic error which is directional and alters mean estimates. TSE combines these two components. The same distinction would also apply under TSMV. In the rest of this article, reference will be to TSE because this is the established and more familiar construct, but it generally would apply to TSMV as well. Conventionally this distinction is used as the first order for distinguishing among types of survey error (Alwin, 2007; Andersen et al., 1979; Groves, 1989; Smith, 1996, 2005). This is usually illustrated by having two separate sets of boxes for each error component with error flows from one set representing bias and from the other variance. There is nothing wrong with this approach, but it introduces into the TSE paradigm a distinction that is different from all other elements of TSE. The other elements essentially relate to from what components of surveys the error originates. For example, TSE categorizes error as coming from sampling, question wording, interviewers, the post-production processes, etc. While the separation of error into bias and variance is important given the statistical differences between these two forms of error, the standard approach of separating these two types of error makes them seem too different from each other from an operational or data collection perspective. It would be better to think that components of a survey can encourage or discourage error and that this error could be bias, variance, or, most commonly, a combination of both.

Figure 1
TSE: Total error

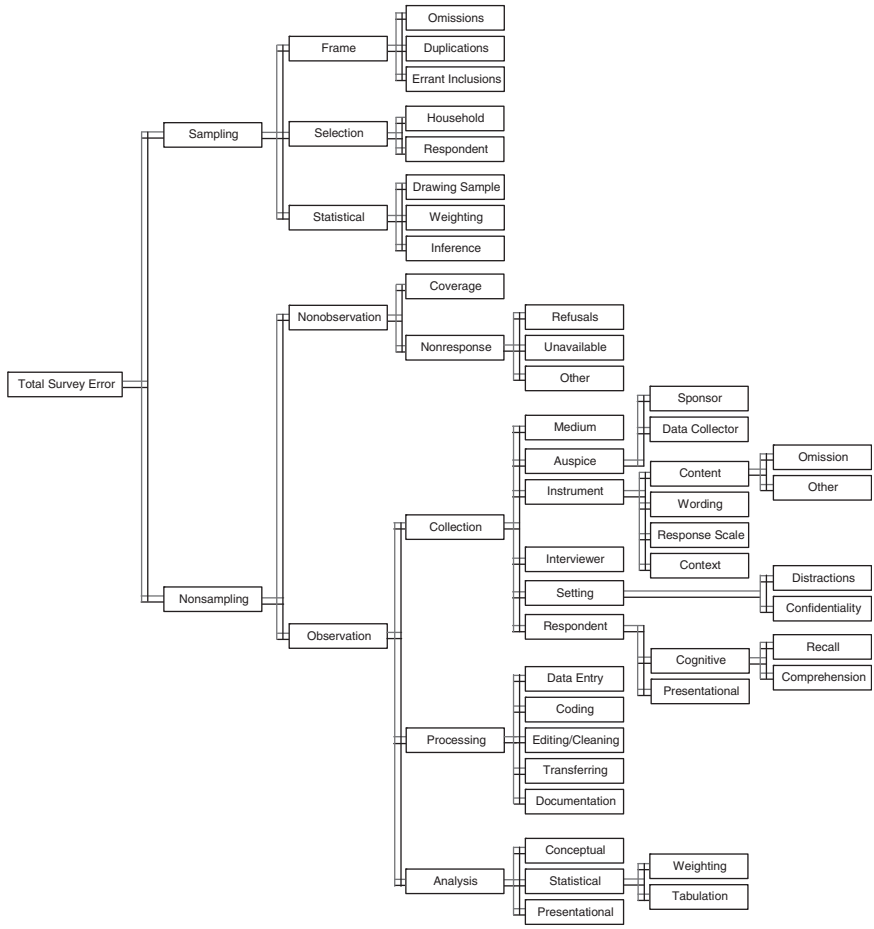


Figure 1 illustrates an alternative way of illustrating TSE with one set of boxes for each error component and two tandem paths from each component, one representing bias and the other variance. This maintains the important bias/variance distinction, but shows each as flowing along with the other from each component. It also has the advantage of eliminating the need to have duplicate boxes illustrating each component twice (e.g., as in Groves, 1989; Smith, 2005).

The "T" in TSE Stands for "Total"

TSE models should be as comprehensive as possible. After all, the modifier "total" does promise completeness. This is difficult first because surveys are

complex instruments with many components and second because there are various alternative ways of categorizing error. It is noteworthy that every major description of TSE from Deming (1944), through Hansen, Hurwitz, and Madow (1953), Kish (1965), Brown (1967), Groves (1989), Smith (1996), Biemer and Lyberg (2003), and Alwin (2007) to this rendering has produced a different taxonomy with some unique elements. Moreover, as Deming (1944) noted about his classification of errors in surveys, "the thirteen factors referred to are not always distinguishable and there are other ways of classifying them [. . .]."

To illustrate the point about alternative and overlapping categorization, consider several examples from Figure 1. Weighting appears in two places. First, it is located near the top, under Sampling. At this point it refers to errors such as miscalculating a weight or failure to create a weight when it is needed. Second, near the bottom, under Analysis Weighting appears again. This refers to such examples as: Failure to use a required weight in analysis, using the wrong weight, or apply the right weight incorrectly. (In each of these latter cases, there is no error in the right weight, but an error in the use of weighting). This separation seems both clear and sensible, but others might wish to accommodate all weighting-related error in a single category.

Second, interviewer error is placed under nonsampling observation error, but one of the things that interviewers usually do is implement sampling protocols such as by completing a Kish table or asking about who had the most recent birthday. Thus, this part of interviewer error might instead be placed under sampling/selection/respondent.

Finally, consider the best place to locate coverage error. Most commonly, it is included as a nonsampling, nonobservation error. But alternatively, it might be possible to consider coverage error as covered by Frame error. That is, if under coverage due to nonobservation is not due to nonresponse, then perhaps all under coverage is a function of errors in the frame or sampling from the frame and these are covered under sampling and there would be nothing left to include under nonobservation.

Next, there are various refinements and extension that can be applied. A few examples will illustrate these. First, consider nonresponse. Figure 1 makes the standard subdivision of nonresponse into refusals, unavailable, and other (e.g., too ill, too handicapped, administration mistakes, lost interviews, etc.). A common alternative subclassification of nonresponse organizes it as nonresponse at the level of the (a) unit, (b) supplement or self-administered module, and (c) item. Moreover, it is possible to combine the two three-category schemes into a matrix classification with nine theoretical categories. Of the nine, it appears that seven combinations actually occur (Figure 2). This figure lists some, but not all, of the situations that account for these different categories of nonresponse. From the TSE perspective, the importance of the

Figure 2
Categorizing nonresponse error

		Level of Nonresponse		
		Unit	**Supplement/ SAQ**	**Item**
Reason for Nonresponse	**Refusal**	Refuse Survey	Refuse Supp./ SAQ	Refuse Question
	Unavailable	Non-Contact	Null	Null
	Other	Illness, Lost Case	Illiterate/ Poor Eyesight	Cognitively Unable

two separate and one combined classifications is whether they help researchers to identify the sources of errors and to develop strategies to ameliorate these errors.

Second, consider that medium is used to cover three aspects of the administration of the survey: (a) The mode of the primary sense being used to ask and retrieve data (e.g., visual, audio, mixed, etc.); (b) the use of technology (computer vs. no computer); and (c) self-administration versus interviewer administered. The main modes of administering the survey are visual, audio, and mixed, combined with computerized versus not computerized and self-administered versus interviewer administered that produces 12 possible combinations. As Figure 3 illustrates, most of these combinations exist, but some are nonexistent or rare. Among the most common for general populations are verbally asked and answered surveys using computers either over the phone (i.e., computer assisted telephone interviews—CATI) or in-person (i.e., computer assisted personal interviews—CAPI). Other popular mediums include class-based, student samples using a noncomputerized, visual medium that is self-administered (i.e., a hardcopy, handout questionnaire that is filled out and returned anonymously) and visual, computerized, self-administered surveys over the Internet. Mixed can refer to a survey using 2+ distinct modes (e.g., the GSS using both in-person and telephone interviews), each interview having 2+ modes (e.g., the GSS using CAPI with computer-assisted self-interviews sections), or a combination of the two forms of mixing. The error structure will of course vary across these various mediums (e.g., greater social deniability effects in interviewer-administered surveys; greater error among the hearing impaired in aural administrations).

One important element that is excluded from Figure 1 is the crucial element of overall design involving the articulation of the research

Figure 3
Typology of surveys by medium.

	Computer		No Computer	
	Interviewer	Self-Admin	Interviewer	Self-Admin
Visual	Null	Internet, CASI	Null	Postal, Classroom Handout
Audio	CAPI, CATI	Automated Voice + Voice recognition	PAPI, TI	Null
Mixed	CAPI+ Show Cards, CAPI + CATI	Automated Voice + Touch Tone Response, ACASI	PAPI + Show Cards	TI +Show Cards, TI+ Diary, PAPI+SAQ

Note. ACASI, audio computer-assisted self-interview; CAPI, computer-assisted personal interview; CASI, computer-assisted self-interview; CATI, computer-assisted telephone interview; Null, rare or non-existent; PAPI, paper and pencil interview; SAQ, self-administered questionnaire; TI, telephone interview

question being investigated and how that question should be studied. It is a formative and overarching matter that draws on both substantive knowledge of the subject being studied and methodological expertise about how to collect data. It is not covered in Figure 1 because here it is conceptualized as an external element that precedes and shapes decisions about the components of surveys. That is, it is present in each and every component in that decisions about how to do the study manifest themselves in each component. As an alternative, it may be possible to incorporate elements of design and theoretical conceptualization as an explicit part of the TSE model, but that seems more problematic than treating it as an implicit, overarching facet.

While acknowledging that any classification scheme is imperfect and that sensible alternative schemes exist for the placement of certain components, it is still important that the classification scheme be both logically organized and as close to comprehensive as possible. To be most useful, each and every source of error needs to be conceptually accounted for. If not included, in effect that source of error is being ignored and that means both TSE is being underestimated and that no specific steps will be taken to understand and minimize error from the omitted source. Figure 1 expands upon the usual renderings, but it is not definitive. In particular, it omits useful distinctions that would appear at the far right. For example, it does not: (a) Make explicit that falsified interviews would be part of Interviewer error; (b) include the nonresponse and medium matrices discussed above and shown in Figures 2

and 3; (c) subdivide Context error into (1) placement-related, concerning warming-up and fatigue effects and (2) substantive order effects; (d) fully incorporate the Tourangeau–Rasinski (1988) question-answering scheme (comprehension/retrieval/judgment/answer selection) under Cognitive; (e) distinguish between open-ended and closed-ended Wording effects; and (f) mention what Biemer and Lyberg (2003, p. 39) call using an "information system" which would go under Wording if a question explicitly called for checking a document for an answer or under cognitive if it was something respondents spontaneously did to aid their recall.

Interactions

Interactions are a key component of TSE, but have been under examined in the TSE literature (Groves, 2005; Smith, 2005, 2008). Interactions are crucial to standard, single-survey applications, to panel surveys, and to multisurvey extensions focusing on comparison error.

Interactions in Single Surveys

To keep discussions of the components of TSE focused, descriptions have tended to examine each component separately and in turn. For example, Groves (1987, p. S162) examined measurement error from the interviewer, survey questions, respondents, and mode, but discussed only "the direct effects of these four sources of measurement error but omits mention of their combined effects." As Groves (1987, p. S168), further noted a "problem ignored in most methodological investigations is the existence of relationships among different error sources [. . .]. (T)here is little work examining the relationships between different error sources."

This neglect is probably encouraged by the standard way of illustrating TSE which shows each source of error as an isolated flow. This could wrongly contribute to the idea that the errors occur independently of one another. Nothing could be further from the truth. In fact, there are more often than not close connections and interactions among the various components of errors. This might be illustrated by drawing lines between different components to indicate their interconnection. This would create a dense web of lines that would correctly visually indicate the numerous and complicated ways in which errors are related to other another. But it would be such a cluttered presentation that it would not be informative.

A few well-established examples of the connection between different components of TSE will illustrate the extent and variety of such interactions

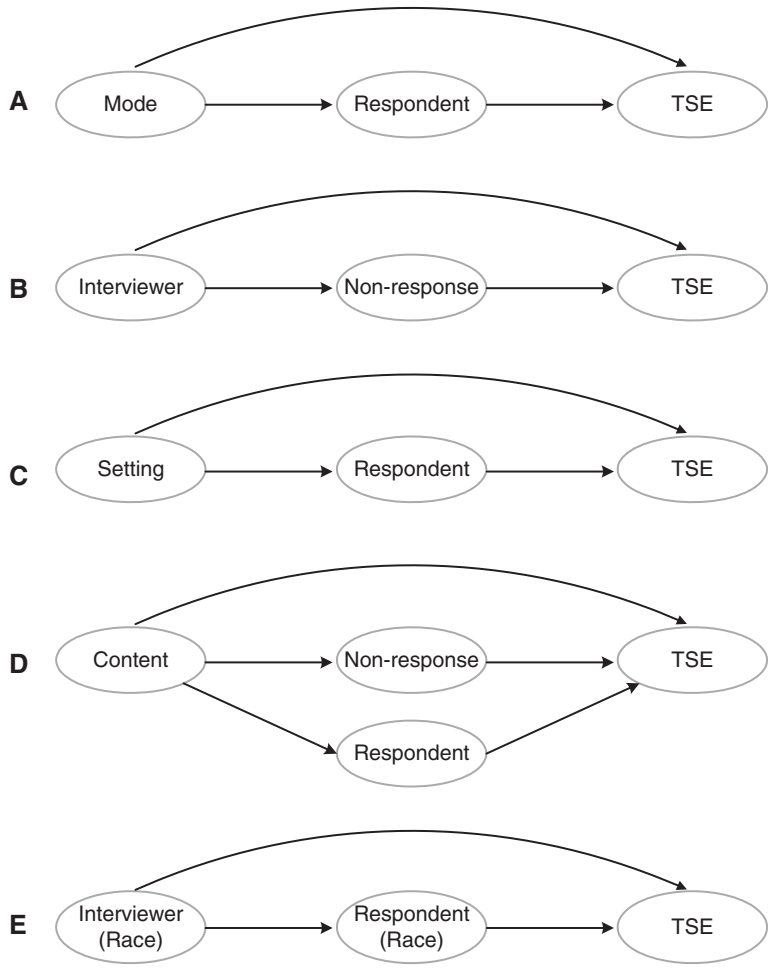
(Groves et al., 2004; Schuman & Presser, 1981; Smith, 2005; Tourangeau, Rips, & Rasinski, 2000):

(a) Self-administration leads to less underreporting of undesirable behaviors, but less education in general and low literacy in particular leads to more error on self-administered questionnaires.
(b) Poor interviewers make more mistakes and get a lower response rate thus contributing to more error from both interviewing error and nonresponse bias.
(c) Respondents with hearing problems would presumably be most error prone in noisy settings.
(d) Content dealing with sensitive topics will contribute to both unit nonresponse and respondent presentational effects (e.g., lying).
(e) Interviewer race interacts with respondent race to affect responses to race-related questions. Under one interpretation this is a presentational bias. It assumes that respondents are lying to please interviewers or hide their racism. Under another interpretation being interviewed by a presumably polite person of another race who is engaged in the scientific activity of sampling their attitudes may truly change their response to race-related questions. Under the first explanation it is clearly a component of TSE, but under the second, while part of TSMV, it is arguable that measurement error is not occurring (Schwarz, Groves, & Schuman, 1998, pp. 153–154).
(f) Less development work/pretesting increases error from both poorer question wordings of individual questions and the greater likelihood that an adequate, multiple-item scale is not constructed.
(g) Using CATI to lower costs over CAPI would both result in undercoverage of households without phones and a lower response rate, but might mean a larger sample size that would reduce sampling variance.

Illustrating these and the many other connections among error components in the overall TSE chart is not practical. Instead, it is more useful to create separate path diagrams that illustrate how particular errors relate to other another (see Figure 4 for illustrative examples).

Interactions might be considered in general to be due to resource constraints or tied to specific interactions between components. That is, having less resources will mean that all components would utilize lower cost and generally more error-prone practices (e.g., less item development, interviewer training, interviewer monitoring, and lower response rate). In other cases, the interactions are because of a direct connection between the components. For example, less interviewer training will lead both to interviewers making more errors and to more errors by respondents since garbled presentation by poor interviewers or such less well-trained interviewers not knowing how to

Figure 4
Interactions with TSE



properly respond to respondent queries will increase error by respondents. As another example, using CAPI means less skip errors and less transfer error, but probably more data-capture error since there are more typos and keystroke errors in CAPI than miscirclings and handwriting errors in PAPI. Moreover, even two random-error components can be correlated. Not in that the specific errors from one source match specific errors from the other source, but in the total amount of random error may be associated. For example, poorly trained interviewers and inattentive or uncommitted respondents may interact with one another to have more TSE than just either poor interviewers or poor respondents would produce alone.

Panel Surveys

TSE also involves interactions that occur across multi-wave panels using intrasubject designs. First, panel surveys have unique sources of error that do no exist in single-wave surveys (Alwin, 2007). Examples include conditioning error in which the experience of being interviewed at an earlier time (e.g., in wave 1) influences how questions are understood and answered in subsequent times (e.g., in wave 2). Additionally, there is false-consistency error in which a person recalls his response to a question in an earlier round and repeats that response to appear consistent even if it does not now and/or did not at the earlier time reflect the true value. Also, there is autocorrelation or autoregressive error (Jung, 2005). Finally, there is wrong-respondent error in which the wrong person is "reinterviewed" (Smith, 2009a).

Second, for a measure of individual-level change to be accurate it would ideally be error free at each wave. Short of the holy grail of error-free data, one would want each panel wave to have the same error structure so that any observed change would only reflect true change. Least desirable would be variable error in which the error structure changed with wave so that any observed change across panel round would be either a function of variable error or an inseparable combination of measurement error and true change.

TSE in Multiple Surveys

TSE is traditionally applied to a particular survey, that is all measurement components are considered, but only in reference to a single study.[1] However, survey research often involves comparing two or more studies. The TSE paradigm can and should be extended to cover more than one study (Munck, 1991; Scherpenzeel and Saris, 1997; Stoop, 2007). When two or more studies are being compared, one needs to consider the interaction across studies of the error structures.

The multiple surveys can differ on several dimensions: (a) Data collector or what organizations conduced the survey, (b) time or when the data collections occurred, and (c) target population or from whom the data are collected. The first dimension is conventionally referred to as house effects (Smith, 1978, 1982). The second dimension is addressed in methodological considerations of studying societal change (Smith, 2006a). The third is most frequently covered by discussions of cross-cultural and cross-national surveys (Smith, 2002, 2004, 2010), but it would also include comparisons between such different target populations as adults and children or residents of households and the institutionalized.

[1]For example, there is no cross-cultural/cross-national discussion of TSE or data quality comparability in Andersen et al., 1979; Biemer & Lyberg, 2003; Groves, 1991; Groves et al., 2004; Smith, 2005.

The greater the difference on any of the dimensions between two surveys, the more problematic is comparability. Thus, comparability would be most suspect the more dissimilar the data collectors were, the greater the time between the surveys, and the larger the difference across the target populations. In terms of target population, differences are greater when there is little or no overlap between the target populations and when the nonoverlapping target populations have less in common (e.g., different languages, cultures, structures, etc.). Additionally, these three dimensions may also occur in combinations. For example, two surveys may differ in terms of both time and target population. Cross-national surveys are especially challenging because they naturally involve both differences across data collectors and the maximum difference across target populations—intersocietal. Data in different countries are almost always collected by different organizations and field staff, and house effects are to be expected. In general, intersocietal differences represent the largest difference across target populations, usually differing on language, culture, and structure and often differing on aspects specifically related to conducting surveys (e.g., privacy laws, social-desirability norms, civil liberties, and survey climate).

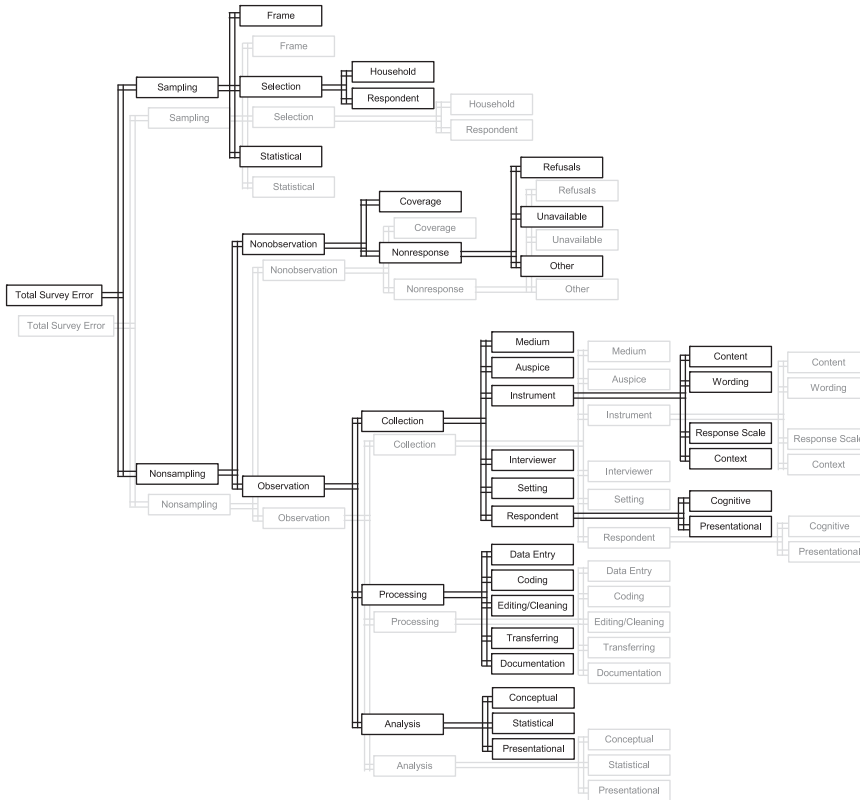
The TSE paradigm is a valuable approach for comparative studies for several reasons. First, it is a blueprint for designing studies. Each component of error can be considered with the object of minimizing comparison error. Second, it is a guide for evaluating error after the surveys have been conducted. One can go through each component and assess the level and comparability of the error structures. Third, it goes beyond examining the separate components of error and provides a framework for the combining of the individual error components into their overall sum. Finally, by considering error as an interaction across surveys, it establishes the basis for a statistical model for the handling of error across surveys. As Figure 5 indicates, each component is measured in each survey, and across each component there is a potential interaction in the error structures.

The interaction in measurement error across surveys leads to what Weisberg (2005) refers to as "equivalence problems" or "comparability effects" or what is referred to here as "comparison error." One can think of such comparison error as occurring both for each component and in the aggregate across all components. For example, errors due to mistranslations are a comparison error that is an interaction between the question wording components of each study.

Functional Equivalence

Functional equivalence or measurement comparability is the standard for research using two or more studies. There are two distinct, but related,

Figure 5
TSE: Comparison error



ways of looking at this goal: (a) From the design and execution perspective and (b) from the measurement error perspective.

From the design and execution perspective the goal is to have surveys designed with similar features (e.g., target population, content, and interviewer training) and carried out to a similar (and hopefully high) level of attainment. That is, they need to be designed to do the same thing and those intentions need to be successfully achieved. Similar designs and procedures alone are not enough, however, to achieve comparability. The level of error is a function both of a survey's design features and the degree to which the protocols are actually realized. Realization will depend on diligence and supervision in general and specific quality-control procedures in particular. If the "proof of the pudding is in the eating," the proof of survey data quality is in the execution of the protocols and the confirmation of the quality of the collected data.

From the TSE perspective the goal is to eliminate differences due to error so that all differences across surveys can be interpreted as reflecting variation

in the true values and not variation due to measurement differences. Or, in other words, to achieve zero comparison error. The TSE paradigm assists this process at the same time by both breaking error down into all of its components and providing a framework for the integration of all components into a comprehensive whole.

Eliminating all error across surveys is of course theoretically and practically impossible. Some types of error such as sampling variance are unavoidable parts of all sample surveys. Other error such as nonresponse bias can be theoretically eliminated, but realistically can only be minimized. A more achievable alternative would be to reduce all error components to their minimal practical level so that error does not overwhelm true variance.

Another desirable goal would be to make error components comparable across studies. This goal can be promoted by adopting comparable study designs and data collection protocols. It is also facilitated by the fact that some error components are quite common and similar across studies and countries. For example, virtually all full-probability surveys underrepresent men and residents of large cities, and while these biases are obviously sources of undesirable error, they are less likely to be sources of comparison error because of their ubiquity. But often error structures are different. For example, sample frames differ in both their nature and quality across countries, and under coverage and other sample-frame errors often will vary across cross-national surveys.

Alternatively, when the error structures are different across surveys, the goal would be to estimate the survey-specific error and to adjust each survey to minimize the error and thus the adjusted comparisons would largely reflect differences in true values rather than differences in TSE.

In general, if the study-design features are equivalent and study fulfillment is similar, one might expect component errors to be comparable and by extension TSE to be on a par across surveys. But while this is a plausible assumption that would often be correct, it cannot be taken as a given. True variation can interact with measurement error to create comparison error. For example, asking about drinking alcohol is not an especially sensitive topic in most European societies, but would be so in conservative Muslim countries. As a result, social-desirability bias concerning alcohol consumption would likely be much greater in the later than the former.

The aim of equivalence in study design and achievement does not mean that procedures need to be identical. For example, having 100% valid interviews would be the goal of most surveys. This goal in general would be promoted by vetting interviewers hired to do the survey, interviewer training on sampling procedures and research ethics, and the supervision/monitoring of interviewers during data collection. In addition, various case-verification procedures are usually employed. In face-to-face surveys in

the United States, the common procedure is to randomly recontact a portion of each interviewer's cases and confirm that an interview had taken place. In other countries, especially in developing countries, interviewers are often sent out in teams with a supervisor accompanying the cadre of interviewers and confirming their work as it occurs. In Germany the Allensbach Institute has not wanted to record the name and contact information of respondents, so verification interviews were not a possibility. It instead developed special techniques to internally validate interviews. One technique was to have the respondents write out responses to an open-ended question. The handwriting could then be examined to see if the interviewer was filling out fake interviews. Another procedure was to have a factual question asking about some obscure matter that almost no one would know and then at a later point in the interview include a second question that in effect supplied the correct answer to the difficult knowledge item. In a real interview, respondents would receive the tip too late to assist them in answering the knowledge item. But an interviewer making up interviews would be aware of the correct answer and would presumably sometimes use that to give a correct response to the knowledge item.

New validation techniques have been developed as CAPI surveys have become widespread in face-to-face, household interviews. One technique is to use the time stamps on the laptops to identify interviews being done much faster than average and/or too close in time between interviews. Another procedure uses computer audio-recorded interviewing (CARI). CARI is used for various substantive reasons such as to more fully record responses to open-ended questions and to allow detailed analysis of speech patterns (Smith and Sokolowski, forthcoming). CARI can also be used to monitor interviewers by checking if questions are being read as scripted and to validate that an interview with a respondent is actually being conducted. CARI, however, cannot readily verify that the interview was conducted with the correct respondent.

As the above examples attest, validation procedures can vary notably across organizations and surveys. This variation is not problematic to the extent that the same outcome of eliminating faked interviews is achieved.[2] But if some techniques are less effective than others, then comparison error will occur in part because of these differences.

In brief, both within single surveys and across multiple surveys, one must continually be alert to the interaction of error components in general and to the problem of comparison error in particular.

[2]Kish (1994) makes a similar observation about probability samples using different sample frame, but still representing equivalent target populations.

Minimizing TSE

TSE can be reduced by applying the best, currently recognized, gold standards and by carrying out research to determine what the best practices are and/or what improved methods can be developed. Of course, it is often not merely a matter of knowledge. The design of surveys is not constrained by only substantive and methodological expertise. There are always practical constraints such as the lack of resources. As Groves (1989) has noted, in surveys there is a trade-off between error and costs. As one survey firm had as its pseudo-model, "quality, speed, price: pick two" (Smith, 1995). More resources (i.e., higher costs) generally reduce most forms of errors. With a given level of resources one wants to make the optimal design choices to minimize TSE. One needs to decide how much to spend on each element such as sample size, response rate, item development and pretesting, interviewer training, data cleaning, etc. The goal in general would be to spend the available resources in a manner that would minimize TSE and to make tradeoffs between different allocations depending on which were more cost-effective in reducing TSE. Unfortunately, one usually has little precise information on the level of error reduction that a given expenditure would obtain or the net error reduction that would be achieved if one component was traded off for another. For example, we know that reducing sample size will increase sampling variance and that increasing interviewer training and monitoring will reduce interviewer error, but rarely know if a particular tradeoff between the two would result in a net reduction in total survey error.

Moreover, information is minimal on the expense involved in various designs, and the cost impact of various study designs can usually only be roughly estimated. While many changes in survey design can be budgeted to a reasonably accurate degree (e.g., the cost of a larger sample size, longer interviewer training, dual coding of all open-ended questions, and more pretesting), the resulting changes in error are rarely known.

Minimizing error is not only a function of design but also depends on execution. Some surveys may be poorly designed, but even more are reasonably designed, but not adequately conducted. Quality control procedures are needed to ensure that surveys are actually conducted as planned.

While each of the components is a source of error, some components should lead to a net reduction in error. For example, while data editing and cleaning may create new errors due to a wrong correction and may fail to correct an existing error, when done properly editing and coding should reduce TSE. Interviewer validation and instrument development via pretesting are other components that should lessen TSE.

Of course, the complexity of doing surveys and the lack of information about many error components means that minimizing TSE is not easy and that often design decisions need to be based on imperfect and incomplete

information. Yet despite these serious impediments, TSE provides a framework for guiding design and execution.

TSE at Individual and Aggregate Levels

Error exists at both the individual/case level and at the aggregate/sample level. At the individual level an error occurs when a wrong (nontrue value) response is recorded for a particular question for a particular respondent (e.g., age is coded as 35 instead of correctly as 53). Error at the aggregate level of course includes the sum of the individual errors, plus sampling and nonobservation error. That is, it contains observational error, plus error due to nonobservation (e.g., nonresponse bias) and errant or unbalanced representation (e.g., from chance oversampling/undersampling of any variable). In this sense TSE is more than the sum of the individual (observed) errors.

But in another sense individual errors are greater than the total observed aggregate error because errors can off-set each other. In the age example above, a second error in which someone 35 was wrongly recorded as 53 would in the aggregate cancel out the other age miscode and the age distribution would be correct. Of course that does not mean that the data are error free. In this example, there are two fairly large errors in age that would distort correlations with this variable.

However, at the individual level two errors may make a right. Say a person miscounts and reports he visited the hospital six times during the last 12 months rather than the correct seven times. If the interviewer mistakenly hits the seven key rather than the six key, then her error cancels out his error and the final data have the correct value. Of course two wrongs can also expand on an error rather than cancel it out. In this example, an interviewer might be just as likely to miskey a five as a seven and this would of course increase the initial error from one to two rather than eliminating it.

Likewise, error exists for individual measures or specific statistics of interest and for surveys as a whole. TSE is item/indicator specific because the error structure is unique to each measure. This is most obvious when error from context and question wording is considered. But error is also shared across items. Sample size and overall study design have similar impacts across all items in a survey (especially when there are no clustering and item-level design effects) and many other components (e.g., interviewer training, case validation, and documentation protocols) tend to have similar influences across specific measures.

TSE as Absolute versus Relative

In general, TSE is thought to be absolute since it represents the difference between the true value and the measured value. But TSE also has a relative or

situational aspect to it. TSE for the same data may differ according to how it is used. For example, if a survey mismeasured the ethnicity of many white respondents, but correctly classified the race of all whites and blacks, then an analysis by blacks versus non-blacks would have zero racial classification error, but an analysis that looked at various white ethno-racial categories would have error. Also, what measure of a construct would have lower TSE would often depend on the purpose for which a variable was to be analyzed. For example, asking about attending religious services during the last 7 days (Did you, yourself, happen to attend church or synagogue in the last seven days?) is better for establishing the % of the population attending religious services in a given week than asking about how often one attends religious services (How often do you attend religious services?). But the last-7-days question is subject to considerable seasonal variation (e.g., did Easter occur during the last week?) and is thus more problematic if one wants to cover time in general. Moreover, it is a poorer measure of how attending religious services influences other attitudes and behaviors. The last-7-days question is distorted by many transitory deviations from ones standard pattern of attending religious services. Dedicated congregants may have missed a week due to illness, traveling, etc., while even the least religious may have attended a wedding, requiem mass, or other special religious event. In effect, for various correlational analysis the last-7-days measure has more random error and as a result lower correlations with other variables than the often-attend measure which is more likely to capture a person's typical behavior.

Summary

TSE is a very valuable paradigm for describing and improving surveys, but it can be improved. First, either TSE needs to be limited to covering just instances of differences between true and measured values or TSE should be rechristened as TSMV if other forms of measurement-related variation are to be included. Second, the TSE/TSMV typology needs to be as detailed and comprehensive as possible. A single, rigid taxonomy of errors is not needed, but all schemes need to account for the whole range of error sources. In particular, it is important to consider how best to incorporate bias and variance and how to include overall study design and conceptualization of the research question into TSE. Of course, as a practical matter information on the error from specific components may not be available, but the typology needs to be exhaustive and one must at least be aware of all error components even when they are not well measured. Third, TSE needs to be thought of as heavily involving the interaction of error components and the concept of comparison error should be used to extend TSE to cover multiple surveys including trend analysis, comparative studies, and longitudinal panels. In effect, two extensive,

but separate, survey research traditions should be brought together. The TSE paradigm and the related data quality perspective have established a conceptually and empirically rigorous and advanced way of handling errors in surveys. The cross-cultural/cross-national concern about functional equivalence and comparability augmented by other literatures that deal with comparability (e.g., those on house effects and the methodology of studying societal change) provides a valuable emphasis on multisurvey comparisons and an useful framework for identifying comparison error. Together by treating comparison error as an interaction between the error in each survey, the possibility of fully modeling and adjusting for different error structures emerges. Of course separating measurement error from true variance is never an easy process. If differences are found across countries, one needs to determine if they originate from true differences, different measurement error, or a combination of the two. For example, if an association is lower in country A than country B, is that due to more random noise in country A which attenuates the relationship or is the true association actually weaker in country A than in country B? Sorting this out is assisted by applying the TSE perspective. Fourth, the minimizing of TSE is an important goal in survey research, and the TSE paradigm can be used as both an applied application and a research agenda to achieve that goal. But minimizing error is a function of both applying the best science and having sufficient resources and what is known about the tradeoffs of costs and errors remains very limited. Finally, TSE has both individual and aggregate components and an absolute and situational aspect and the role of each of these needs to be kept in mind. In sum, TSE is a powerful paradigm for organizing and improving survey research. But that paradigm itself needs refinement and extension to achieve its maximum potential.

References

Alwin, D. (2007). *Margins of error: A study of reliability in survey measurement*. New York: John Wiley & Sons.

Andersen, R. Kapser, J., Frankel, M. R., and Associates *et al.* (1979). *Total survey error: Applications to improve health surveys*. San Francisco: Jossey-Bass.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. New York: John Wiley & Sons.

Brown, R. V. (1967). Evaluation of total survey error. *The Statistician, 17*, 335–356.

Cantril, H. (1940). America faces the war: A study in public opinion. *Public Opinion Quarterly, 4*, 387–407.

Cantril, H. (1947). *Gauging public opinion*. Princeton, NJ: Princeton University Press.

Deming, W. E. (1944). On errors in surveys. *American Sociological Review, 9*, 359–369.

Groves, R. M. (1987). Research on survey data quality. *Public Opinion Quarterly, 51*, S156–S172.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley & Sons.

Groves, R. M. (1991). Measure error across disciplines. In P. P. Biemer (Ed.), *Measurement errors in surveys* (pp. 1–29). New York: John Wiley & Sons.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R., *et al.* (2004). *Survey methodology*. New York: John Wiley & Sons.

Groves, R. M. (2005, March). *Total survey error: Past, present, and future.* Paper presented to the International Total Survey Error Workshop, Washington, DC.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.

Jung, H. (2005). *A test for autocorrelation in dynamic panel data models.* Unpublished manuscript, Institute of Economic Research, Hitotsubashi University.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review, 62*, 167–186.

Lessler, J. (1984). Measurement error in surveys. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (pp. 405–440). New York: Russell Sage.

Munck, I. M. E. (1991). Path analysis of cross-national data taking measurement errors into account. In P. P. Biemer (Ed.), *Measurement errors in surveys* (pp. 599–616). New York: John Wiley & Sons.

Scherpenzeel, A. C., & Saris, W. (1997). The validity and reliability of survey questions. *Sociological Methods and Research, 25*, 341–383.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D.T. Gilbert, S. Fiske & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 143-179). Boston: McGraw-Hill.

Smith, T. W. (1978). In search of house effects: A comparison of responses to various questions by different survey organizations. *Public Opinion Quarterly, 42*, 443–463.

Smith, T. W. (1982). House effects and the reproducibility of survey measurements: A comparison of the 1980 general social survey and the 1980 American national election Study. *Public Opinion Quarterly, 46*, 54–68.

Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter: An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly, 51*, 75–83.

Smith, T. W. (1989). Random probes of GSS questions. *International Journal of Public Opinion Research, 1*, 305–325.

Smith, T. W. (1991). Ballot position: An analysis of context effects related to rotation design. In P. E. Biemer R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman, *et al.* (Eds.), *Measurement errors in surveys* (pp. 57–72). New York: John Wiley & Sons.

Smith, T. W. (1995). Trends in non-response rates. *International Journal of Public Opinion Research, 7*, 157–171.

Smith, T. W. (1996, March). *Total survey error: The art and science of survey design.* Inaugural address of the Howard Beers lecture series. University of Kentucky, Lexington.

Smith, T. W. (2002). Developing comparable questions in cross-national surveys. In J. Harkness, F. van de Vijver & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 69–92). London: Wiley Europe.

Smith, T. W. (2004). Developing and evaluating cross-national survey instruments. In S. Presser J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin & J. Martin, *et al.* (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 431–452). New York: John Wiley & Sons.

Smith, T. W. (2005). Total survey error. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 857–862). New York: Academic Press.

Smith, T. W. (2006a). *Formulating the laws of studying societal change*. Chicago: NORC.

Smith, T. W. (2006b). *Wording effects on national spending priority items across time, 1973–2004* (GSS Methodological Report No. 107). NORC, University of Chicago.

Smith, T. W. (2008, September). *Applying the total survey error paradigm to cross-national research.* Paper presented to the 33th Conference on Logic and Methodology in Sociology of the Research Committee, Naples, Italy.

Smith, T. W. (2009a). *2006–2008 General social survey panel validation* (No. 113). GSS Methodology. GSS Methodology. Chicago: NORC.

Smith, T. W. (2009b). Trends in national spending priorities, 1973–2008. Unpublished Report. NORC, University of Chicago.

Smith, T. W. (2010). Surveying across nations and cultures. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed., pp. 733–764) San Diego: Elsevier.

Smith, T. W., & Sokolowski, J. (forthcoming). Using audio-visuals in surveys. In S. Hesse-Biber (Ed.), *The handbook of emergent technologies in social research*. Oxford: Oxford University Press.

Stoop, I. (2007, December). *Survey quality and comparative studies.* Paper presented to ESS Train, Mannheim, Germany.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299–314.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

Biographical Note

Tom W. Smith is Director of the Center for the Study of Politics and Society at NORC/University of Chicago and Principal Investigator of the General Social Survey. He is also co-founder of the International Social Survey Program.