Introducing the Total Survey Error framework.

DARREN W PENNAY

Introduction

This paper leads off the two conference sessions dedicated to Measurement and Other Errors in Survey Research (with a focus on Total Survey Error). It is the first paper because it introduces the concept of Total Survey Error (TSE) and thereby provides a conceptual framework for the other papers in this stream.

Participants in these two sessions will gain an understanding of:

- The TSE framework;
- The different sources of error;
- The usefulness of the TSE perspective in informing approaches to survey design and driving improvements; and
- practical examples of how to reduce TSE.

In order to set the scene this paper provides a top-level introduction to the TSE framework. The format is as follows:

- Placing TSE within a Total Survey Quality framework;
- The survey lifecycle from a design perspective;
- The survey lifecycle from a TSE perspective including a brief explanation of the various errors;
- Measuring TSE;
- Examples of typical TSE trade-offs; and
- A summary of the strengths and weaknesses of the TSE approach.

DARREN PENNAY is the Chief Executive Officer and Head of Research at the Social Research Centre. He is also an Adjunct Senior Research Fellow at the Australian Centre for Applied Social Research Methods (AusCen) at the Australian National University and an Adjunct Professor with the Institute for Social Science Research (ISSR) at the University of Queensland.

Total Survey Error within a Total Survey Quality framework

Total Survey Error refers to the "accumulation of all errors that may arise in the design, collection, processing and analysis of survey data" (Biemer, 2010). The Total Survey Error paradigm relates to making survey design decisions, and sometimes trade-offs, so that resources are allocated in such a way so as to reduce TSE for key estimates. As such, TSE is about *optimising* your survey design within given resource constraints – this is sometimes referred to as 'fit for purpose' design.

The TSE paradigm is part of a much broader concept of *total survey quality*. Whereas TSE is primarily focussed on the deviation of a survey response from its underlying true population value, the total survey quality framework introduces other dimensions of importance to data users such as credibility, comparability, timeliness, etc. If these other dimensions are ignored and the sole focus of the researcher is on minimising TSE the result could be data that are released behind schedule, difficult and costly to access and inadequately documented.

Today, many national statistical agencies, including the Australian Bureau of Statistics (Australian Bureau of Statistics, May 2009), have a total survey quality framework which guides their overall approach to survey research. Minimising Total Survey Error is just one part of this framework. Most Total Survey Quality frameworks have dimensions similar to those outlined in Figure 1.

Figure 1: Common dimensions of a Survey Quality Framework

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

Source: (Biemer, 2010)

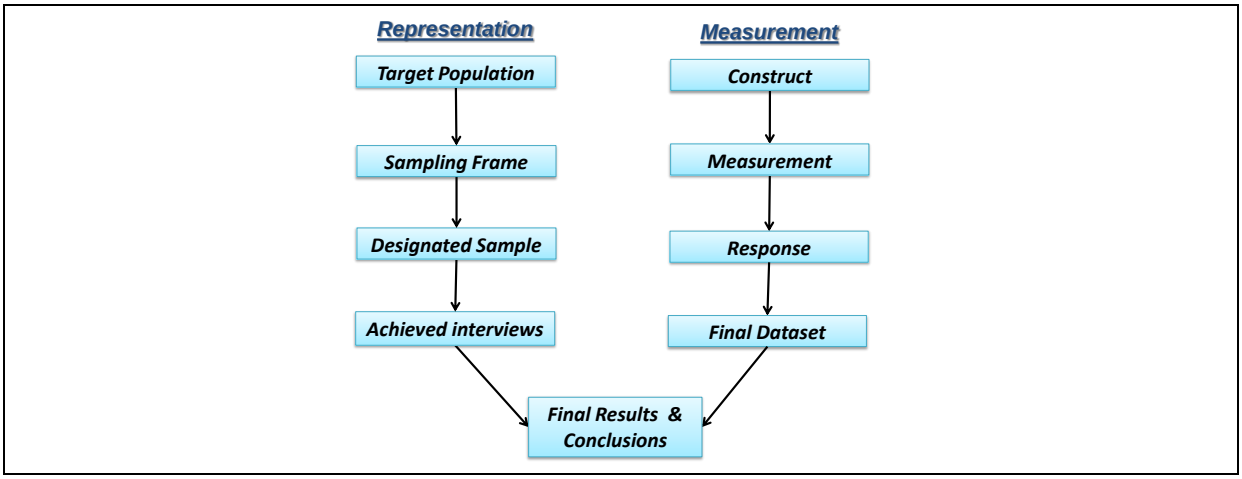
The survey lifecycle from a design perspective

To understand Total Survey Error one first has to understand the survey lifecycle from a design perspective. This is summarised in Figure 2.

On the representation side of the model we have:

- **Target population**: This is the population about which we want to make inferences. Examples include households in Victoria, persons aged 18 years and over, customers of a particular product / service, etc.;
- **Sampling frame**: The list that contains those members of the target population with a chance of being selected in the survey. In its most simple form the sample frame is a list of all units in the target population but often times this is not the case. Sample frames can be incomplete or imperfectly linked to the target population;
- **Designated sample**: The sample selected from the sampling frame. Sometimes called the sample pool.
- **Achieved interviews**: Survey respondents.

Figure 2: The survey lifecycle from a design perspective



On the measurement side of the model we have:

- **Construct**: The item / element we are interesting in measuring (e.g. unemployment, attitudes to a service encounter, crime victimisation, etc.);
- **Measurement**: In survey research the method of measurement is usually via survey questions;

- **Response**: The data produced by the survey questions;
- **Final dataset**: Responses are usually edited, transformed and amalgamated to form a final data set;
- **Results and conclusions**: The two halves of the model (representation and measurement) combine to enable inferences to be made about the target population with respect to the items of interest. These are usually drawn together in the form of results and conclusions.

The survey lifecycle from a Total Survey Error perspective

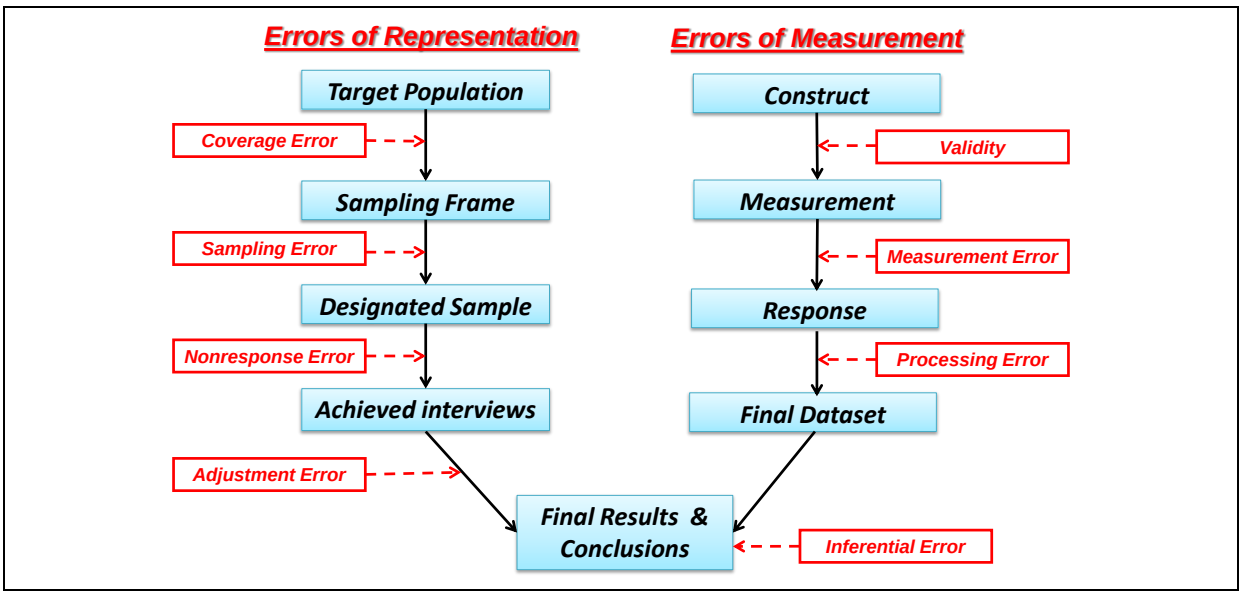
Figure 3 (below) adds 'errors' to each stage of the survey lifecycle. The role of a survey researcher when designing a survey is, at the very least, to be cognisant of these errors and transparent about them. Preferably, however, the researcher will design her or his survey in such a way as to minimise these errors to the extent possible within existing constraints. This often involves making informed trade-offs.

Each of these errors associated with the TSE framework is briefly described, along with relevant examples. We start by looking at errors of representation, sometimes referred to as errors of non-observation.

Coverage and coverage error: Non-coverage occurs when members (units / elements) of the Target Population are not included in the sample frame. This could occur, for example, when there are omissions / exclusions from the administrative list provided to undertake a client survey. Non-coverage error occurs when the members (units / elements) not covered by the sample frame differ on key measures from those included in the sample frame in non-ignorable ways. This leads to biased survey estimates.

In my experience it is quite common for researchers to accept at face value that the sampling frame they have is 'adequate' for their purposes. In reality, however, this is often not the case.

Figure 3: The survey lifecycle from a TSE perspective



Sampling error: Is the difference between an estimate derived from a sample survey and the "true value" that would be obtained if the whole survey population were enumerated. Sampling error has two components bias and variance (Lavrakas, 2008). Bias is a directional source of error and indicates that there is a systemic error in the sample that is present no matter how many times the sample is drawn or a survey conducted whereas variance (or imprecision) is a non-directional source of error (e.g. something which increases the size of the confidence interval of a survey estimate such as the approach taken to sample design / weighting, etc.

Questions to consider when contemplating the sampling approach to be adopted for any particular survey include, but are not limited to:

- What degree of statistical precision, if any, is required?;
- How many units/elements will be chosen to create the Designated Sample?;
- Are the basic requirements for a probability sample met? (e.g. random for each unit/element, non-zero probability of selection for each unit/element and known probability of selection for each unit/element);
- What sample design will be used (e.g. multistage sample, clustering, stratification, etc.).

An important consideration is that if the sample deign has the features of a probability sample (i.e. a randomly selected sample with a known, non-zero chance

of selection for each unit / element of the target population) then sampling error can be quantified. With sampling error quantified, confidence intervals can be calculated. This enables inferences to be made about the target population with some degree of confidence about the likely margin of error of those estimates.

When a nonprobability approach to sampling is being used, an example being the convenience-based sampling methods to recruit opt-in online panels, then sampling errors cannot be calculated and "researchers should avoid [such methods] when one of the research objectives is to accurately estimate population values." (Baker, et al., 2010).

Nonresponse error: There are two types of nonresponse error, unit nonresponse and item nonresponse:

- **Unit nonresponse**: It is very rare to get a 100% response rate for any survey. As a result we end up with unit nonresponse, that is, not all units within the Designated Sample have responded to the survey. Reasons for unit nonresponse include those relating to noncontact, refusals and being unable to complete the survey. Unit nonresponse isn't a problem of itself but if the non-respondents differ from respondents to a non-ignorable degree on key measures then we have nonresponse bias.

  The landmark study undertaken by the US based Pew Research Centre '*Assessing the Representativeness of Public Opinion Surveys'* (The Pew Research Center, 2012) showed a decline in typical response rates for telephone surveys from 36% in 1997 to 9% in 2012 (AAPOR Response Rate 3). While the study did not conclude that declining response rates are without consequences the overarching conclusion drawn from the study was that "telephone surveys that include landlines and cell phones and are weighted to match the demographic composition of the population continue to provide accurate data on most political, social and economic measures." In other words, nonresponse is not synonymous with nonresponse bias.

- **Item nonresponse**: It is also very rare for every respondent to provide data on every measure (e.g. don't know / refused / skipped). The missing items are referred to as item nonresponse. When part of the final sample that does not provide substantive data for a given measure differs in non-ignorable

ways on this measure from those in the final sample that do provide data, then Item Nonresponse Error is said to have occurred. When this happens it typically biases the findings of the study. A good example of such a phenomenon is attempting to collect income data in general population surveys. Approximately one in five respondents typically do not answer such questions and the missing data is, generally speaking, not missing at random with higher income individuals less likely to respond to this question (Yan, et al., 2010).

Adjustment error: It is often the case that the final sample needs to be adjusted for the design effects introduced by the sample design as the unadjusted sample will not be representative of the population for reasons related to Coverage, Sampling, and Nonresponse. **Weighting** is the technique used to 'adjust' the data in order to reduce bias. When we weight our data we are adding error in the form of **variance** (imprecision) to the study's findings. This is because we are introducing a "design effect" (***deff***). The sampling error for any survey must be multiplied by the deff (in doing so an effective sample size is created). The effective sample size is typically some smaller number than the final sample size and thus the confidence intervals for the study are inflated.

Now moving to the measurement errors, sometimes referred to as errors of observation.

Validity: To the extent that the measures we use do not adequately capture the construct of interest then we have an invalid measure – sometimes called a specification error. Researchers are often most concerned about whether a series of items (e.g. a battery of questions) measures the underlying construct of interest – this is the area known as psychometrics (i.e. psychological measurement theory). For example:

- Do the tests we have constructed for students accurately measure the curriculum they have been taught / their year-level mathematical ability?
- Do the elements we have measured really relate to the concept of "satisfaction", "psychological distress", "anxiety and depression", "post traumatic stress", etc.?

Fortunately there are statistical tests to help us measure aspects of validity (e.g. Item Response Theory, Cronbach's alpha, etc.).

Measurement error:

If 'validity' is a high-level error, that is, are we actually measuring what we think we are measuring?, at the next level down there are many more opportunities for Measurement Error to be introduced. Measurement Error can be:

- Questionnaire-related: the product of poor questionnaire design (e.g. order effects);
- Respondent-related: respondents may provide inaccurate answers due misunderstanding a question, a lack of cognitive effort, etc.;
- Interviewer-related: due to poor interviewing technique (e.g. not reading the questions as written , poor probing, not recording answers accurately);
- Mode-related: different modes of data collection can contribute to different types of error. For example, primacy effects are more commonly associated with hard copy self-completion questionnaires whereas recency effects when response options are more common in telephone surveys.

Processing error: The "raw data" that are gathered in a research study typically need to be processed before they can be analysed. This includes:

- Fixing or dropping "bad" data / the treatment of outliers
- Coding raw data (e.g. open-ended verbatims) into other forms
- Imputing missing data
- Deriving new variables
- Appending auxiliary variables.

Each of these processes has the potential to add to the TSE for a particular survey / item within a survey.

Inferential error: The TSE model used in this paper (refer back to Figure 3) contains an additional source of error not included in all TSE frameworks. This is labelled as inferential error and refers to the types of errors that can be introduced to the survey

process at the stage of interpreting the survey findings. Examples of inferential error include:

- Inferring causality when such an inference is not supported by the research design (i.e. not a longitudinal survey or an experimental design);
- Drawing inferences beyond the statistical limits of the design (i.e. not reporting or misreporting statistical significance);
- Drawing population inferences from a nonprobability sample that does not adequately represent the population of interest;
- Using incorrect statistical techniques (e.g. calculating confidence intervals from nonprobability sample such as opt in online panels); and
- Simply drawing the wrong conclusions from the data – perhaps reflecting one's own biases, prejudices, preconceived ideas, preferred outcomes, inferred pressure, desire to please etc.

This is often the area where professional independence, ethics and judgement come into play.

Measuring TSE

Making correct design decisions and correct decisions regarding the allocation of resources with a view to minimising TSE implies having some knowledge about the relative impact of discrete sources of error on the overall TSE for a particular estimate. For example, if nonresponse is thought to be a larger source of error than measurement error arising from interviewing practices then more resources can be allocated to response maximisation and less to interviewer training and monitoring. While it is theoretically possible to measure TSE using approaches such as the Mean Squared Error approach[2] (a topic not further explored in this paper) in practice this is rarely the case as an unbiased estimate if the parameter of interest is needed.

Fortunately, however, detailed knowledge on the costs, errors and methodological effects are not needed for every survey design as many such findings have already been published in the survey research literature and, often times are generalisable to

[2] In it's simplest form Mean Squared Error (MSE) is the mean of the squared differences between the predicted parameter and the observed parameter. From a TSE perspective it is often also useful to think of MSE as the composite of the variance of the estimated parameter and some unknown random bias $Var(\theta^*) + Bias(\theta^*, \theta))^2$ which enables us to say that an unbiased parameter estimate should have the same MSE as the variance of the estimate (Lavrakas (ed.), 2008).

similar studies (Biemer, 2010). For instance, views on the optimal design features for mail surveys are widely held based on the work of Dillman and others (Dillman, et al., 2009). This is also the case for internet surveys (Couper, 2008), (Dillman, et al., 2009).

The optimal design features to be adopted for interviewer administered surveys are less well advanced given the increased variation resulting from the role of interviewers. However, there is still an extensive literature regarding the lessons already learned (e.g. well-designed advance notification letters are commensurate with best practice).

Also, when there is a paucity of relevant information to draw on we can conduct our own experiments aimed at reducing TSE. This is especially useful for those researchers involved in longitudinal surveys or time series (repeat cross sectional) surveys / monitors. The best way to undertake this type of experimentation is to derive some proxy indicators of data quality for our surveys and then measure the impact of alternative designs on data quality. Some practical example of typical TSE trade-offs now follow.

Examples of TSE trade-offs

Coverage error versus measurement error

In December 2011 the Social Research Centre conducted its first dual-frame omnibus survey. This was a subscriber survey that used two sampling frames in order to try and improve the coverage of the target population. A sample frame comprised of randomly generated landline telephone numbers and a sample frame comprised of randomly generated mobile phone numbers. The inclusion of mobile phone numbers overcomes a known source of non-coverage bias associated with landline telephone surveys, this being, the exclusion of the approximately one in five adults residing in households without a landline telephone connection.

While the inclusion of respondents from mobile-only households seems like a sensible step to take in order to reduce a fairly major error of representation, from a TSE perspective we need to know that obtaining interviews from persons via a

mobile phone does not lead to an even greater source of bias by introducing mode-related measurement error.

At face value there seem to be some good reasons to be concerned about the quality of the data obtained from someone interviewed via a mobile phone. These include:

- respondents may choose to participate in a survey while others can hear their responses and as a result may, even inadvertently, censor and/or alter how they respond to questions;
- the sometimes poor audio quality of mobile phone connections;
- noise in the surrounding environment;
- time constraints the respondent is under that cause her/him to rush to complete the interview, and
- engaging in a wide array of other cognitively "distracting" activities while participating in the interview.

Not all of these data quality concerns are limited to mobile phone responses as landline respondents, especially those responding to a survey on a cordless landline, may also be engaging in other activities while being interviewed or be concerned about censoring their answers due to family being present.

So the question from a TSE perspective becomes whether in reducing one sort of error (coverage error) we have actually introduced another sort of error (measurement error) which could, potentially pose more of a problem to the overall quality of our survey data.

Research into this issue was undertaken by noted US survey methodologist Paul Lavrakas (Lavrakas, 2012).

The following hypothesis were formed:

- Data quality will be lower among the group of mobile phone respondents interviewed away from home compared to the groups of mobile phone and landline respondents interviewed at home;
- This difference will remain after controlling for demographic differences among the different groups of respondents;

- Data quality will be lower among the groups of mobile phone and landline cordless respondents that are engaging in “distracting” other activities while being interviewed; and
- This difference will remain after controlling for demographic differences among the different groups of respondents.

And the following variables were constructed as indicators of data quality:

- A count of the number of items overall for which a respondent said “Don’t Know” or “Refused;” i.e., total amount of missing data
- A count of the number of sensitive questions which a respondent refused to answer or claimed to not know
- Two indicators of the variability of answers within an 8-item set using a Likert response scale, used as indicator of “straight-lining”. These being:
    - The tendency for someone to satisfice by simply picking the same answer over and over again within a series of items with the same response options
    - The strength of intercorrelations among a key health item and key demographic variables.

To assist with this experiment the survey also included some questions to measure the number and type of cognitively distracting behaviours a respondent was engaged in while participating in the interview.

For those interested in undertaking TSE experiments within their own survey programs will often have access to or be able to derive similar indicators of data quality.

Without going through the results of this particular study in detail the findings from this methodological research supported the following conclusions:

- There was some, but not great, indication of poorer data quality associated with “mobility” while being interviewed; and
- Extreme Straight-Liners appear more likely with cordless landline and mobile regardless of location.

For this example (evidence not shown), the large reduction in non-coverage error gained as a result of including persons from the mobile phone sample frame more than offset the slight increase in measurement error associated with interviewing over a mobile device. In this case, the overarching conclusion is that TSE was reduced by the inclusion of interviews with persons via the mobile phone.

The relationship between nonresponse and measurement error

Of course, however, not all efforts to reduce TSE might be as self-evident or successful and it can sometimes be the case that some of the design decisions researchers take might increase TSE without us ever really realising that this is the case.

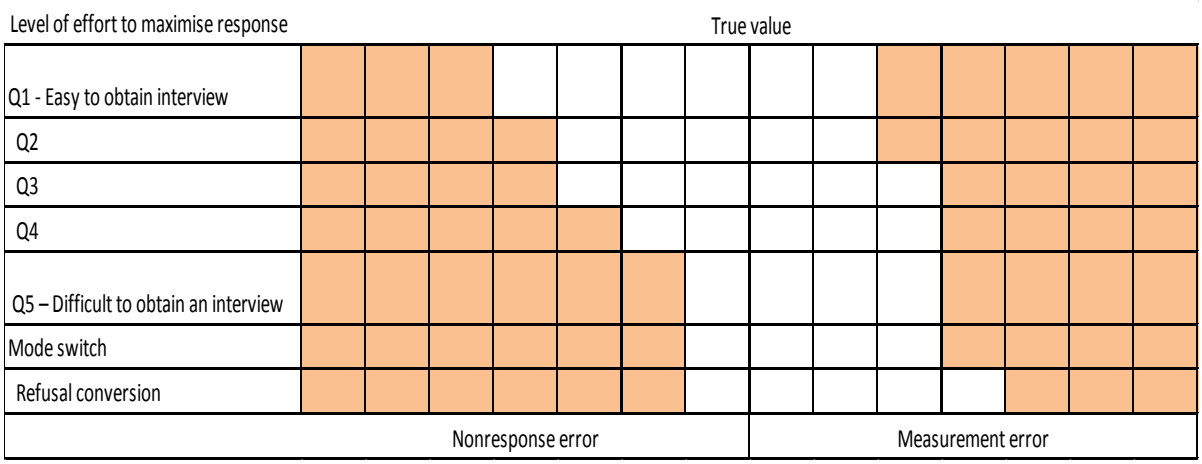
It is quite common for researchers to undertake efforts to increase response rates, thereby most likely reducing errors of representation, without taking into account the possible impact on the measurement side of the TSE model. For example, respondents interviewed as a result of a refusal conversion activity or an extended call regime may be less motivated to put in the effort required to answer questions accurately thereby increasing measurement error. A related dilemma may be attempting to increase response rates by offering multiple modes of data collection and thereby increasing the potential for measurement errors relating to differential mode effects.

By using paradata and deriving data quality indicators for key survey items researchers can start to make informed decisions about which particular combination of survey design features is optimal for any given study.

Figure 4 (next page) is a crude depiction of how the relationship between response maximisation and measurement error sometimes works. This is based on the work undertaken by Kreuter, Muller and Trappman in relation to the German Panel Study "Labour Market and Social Security" as reported in Public Opinion Quarterly (Kreuter, et al., 2010). The German Panel Study primarily adopted a telephone interviewing methodology to enumerate a population of persons in receipt of income benefits. The level of effort to contact sample members over the telephone has been spilt into quintiles with Q1 representing easy to obtain interviews ranging to Q5 – the hardest to obtain respondents over the telephone, requiring 15+ calls. Following on

from telephone call attempts there was a mode switch, in this instance to CAPI, in order to further boost response followed, finally, by refusal conversion interviewing.

Figure 4: Relationship between nonresponse error and measurement error



The relationship depicted in Figure 4 shows the following:

- Nonresponse bias reduces (i.e. we get closer to the true value) up to and including Q5. The increased efforts made to reduce nonresponse error by switching modes and undertaking refusal conversion interviewing, even though such activities increase response rates, do not reduce nonresponse error;
- As efforts to increase response rates progress to mode switching and refusal conversion interviewing, measurement error actually increases;
- In this example, the optimal design would have been not to proceed to mode switching and refusal conversion interviewing in an effort to reduce TSE.

Researchers can conduct their own experiments to inform such design decisions. The benefits to be gained with respect to optimising your research designs are potentially far reaching.

TSE - Summary of strengths and weaknesses

Figure 5 (next page) provides a brief summary of the strengths and weaknesses of TSE as outlined by Groves and Lyberg (Groves & Lyberg, 2010). From my own perspective the TSE framework:

- Provides both a theoretical and practical framework for all aspects of survey design and evaluation;
- Enables researchers to challenge accepted paradigms regarding the primacy of response rates as an indicator of survey quality;
- Helps guide our survey design decisions;
- Is an excellent framework for teaching research students and young researchers about the survey cycle;
- Can be used as an organising framework for proposals and technical reports;
- Is a tool for evaluating our survey designs, helping us make informed decisions and driving continuous improvement, and
- Makes commercial sense from the point of view of research suppliers and research buyers in that optimal research design equates with value for money.

Figure 5: Strengths and weaknesses of TSE

Weaknesses	Strengths
Despite being around since 1944 (Denning) it has not become the dominant paradigm for survey researchers.	Provides a theoretical and practical framework for survey methodologists
Total MSE can rarely be completely measured which makes fully implementing a TSE approach challenging	Decomposition of errors and separation of issues
Survey researchers remain primarily focused on sampling errors	Increases the focus on non-sampling errors
The exclusion of key quality concepts found in overarching total quality frameworks	Increases the focus on achieving optimal design outcomes
	Makes explicit what is otherwise implicit
	Can be adapted for all forms of social, behavioural and market research (including qualitative research).

Concluding remarks

The rest of the papers presented across these two sessions will provide practical examples of errors of representation and errors of measurement and how the researchers involved have attempted to overcome them in order to reduce TSE. Topics covered include:

- the relationship between mode of data collection and measurement error;
- measurement error arising from issued pertaining to construct validity and mode of data collection;
- inferential error;
- coverage error, and
- sampling error.

Acknowledgements:

This paper is based, in part, on a workshop module delivered by Lavrakas and Pennay for the Australian Market and Social Research Society workshop in July, 2013 entitled “Recent developments in Dual-frame RDD surveys”.

References

Australian Bureau of Statistics, May 2009. *ABS Data Quality Framework,* s.l.: s.n.

Baker, R. et al., 2010. AAPOR Report on Online Panels. *Public Opin Q,* 20 October, 74(4), pp. 711-781.

Biemer, P. J., 2010. Total Survey Error: Design, Implementation, and Evaluation. *Public Opin Q,* 74(5), pp. 817-848.

Couper, M. P., 2008. *Designing Effective Web Surveys.* s.l.:Cambridge University Press.

Dillman, D. A., Smyth, J. D. & Christian, L. M., 2009. *Internet, mail, and mixed-mode surveys : the tailored design method.* 3 ed. s.l.:John Wiley & Sons.

Groves, R. M. & Lyberg, L., 2010. Total Survey Error: Past, Present, and Future. *Public Opin Q,* 74(5), pp. 849-879.

Kreuter, F., Muller, G. & Trappmann, M., 2010. Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opin Q,* 74(5), pp. 880-906.

Lavrakas, P. J., 2008. *Encyclopedia of Survey Research Methods.* s.l.:Sage.

Lavrakas, P. J., 2012. *SRC Workshop:Telephone Surveys and the Mobile Phone Only Population.* Melbourne, s.n.

The Pew Research Center, 2012. *Assessing the Representativeness of Public Opinion Surveys,* s.l.: s.n.

Yan, T., Curtin, R. & Jans, M., 2010. Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics,* 26(1), pp. 145-164.