ITSEW 2012
ABSTRACTS OF ALL PAPERS/POSTERS

September 2-4, 2012, Santpoort Noord, The Netherlands

Competing Goals of Responsive Design in a Total Survey Error Framework: Minimization of Cost, Nonresponse Rates, Bias, and Variance

ANDY PEYTCHEV, Research Triangle Institute (RTI), United States

Responsive design is a framework that combines planning, active monitoring, and potential changes to procedures during data collection. The framework does not specify the aims, nor should it attempt to do so. It is imperative, however, that the goals of a responsive design in a particular study are clearly stated as this would allow the identification of the most direct way of achieving them. This paper, and hopefully the ensuing discussion at the workshop, will focus on the past and current objectives of implemented responsive designs, suggest what other important objectives should be considered, and propose methods to combine multiple objectives. In addition, data from a national telephone survey will be used to (1) evaluate the degree of disagreement of the different main objectives, and (2) provide empirical evidence of the impact on survey estimates and survey inference from selecting each approach.

Examples of responsive design from several countries are: to increase response rates by targeting cases with large sampling weights, reduce costs by targeting cases with high response propensities and with low interviewer travel costs, increase the sample size by selecting two adults in each household, reduce potential for nonresponse bias by targeting groups with more variable response propensities based on sample member characteristics, and reduction of nonresponse bias and variance by targeting cases with low response propensities. Both the aims and the operationalizations clearly vary across studies, and in some cases, lead to the exactly opposite approaches (e.g., targeting cases with the highest or cases with the lowest response propensities).
The aims can be classified into four major types:

1. Maximize response rates;
2. Reduce cost (per completed interview, or analogously, increase the number of interviews given a fixed budget);
3. Reduce nonresponse bias; and,
4. Reduce variances (increase effective sample size).

The third and fourth type can be further broken out by source of survey error, but for ease of disposition, that is left for the focus of another study. These two types can also be combined into a single objective through mean square error or some other function of bias and variance, but the four types above aim to be parsimonious; any combination of the above is possible.

It is important to restate that these different aims can be at odds with each other. In a particular study, maximizing the response rate and minimizing cost can be attained by targeting likely respondents in an overrepresented sample stratum (e.g., high propensity cases in the stratum with listed numbers in a random-digit-dial telephone survey). That approach may fail to reduce bias (interviewing “more of the same”) and fail to make estimates more precise (increase weight variability among the final pool of respondents).

At the most fundamental level, the choice is between optimizing the design for achieving higher response rates, reducing the cost per interview, and reducing error in survey estimates (in whatever way that is defined). The aims can then be made more specific, and if multiple aims are needed, compromises need to be made (e.g., reducing cost and bias may require compromises in reducing each). The choice needs to be driven by the ability to best meet the *study* objectives, such as attaining a target level of precision or a mandated response rate. Other important considerations are, arguably, often overlooked despite their importance. If little information is available on respondents and nonrespondents, for example, then targeting cases to reduce nonresponse bias is unlikely to be fruitful. The reduction of variances, however, can still be very beneficial, as the goal then would be to increase the effective samples size (and as a consequence, decrease the mean square error of survey estimates).

To help engage in a constructive discussion that may help advance work in responsive design, pursuit of different aims will be described from past studies and from simulated on data from a two-phase telephone survey that implements different survey protocols in each phase. This is achieved in the simulation through: (1) oversampling of cases with high response propensities, (2) oversampling cases with low response propensities, (3) oversampling cases with large selection weights, and (4) a simple random sample of nonrespondents at the end of the first phase of data collection.

Evidence of a Coverage-Nonresponse Trade-off

STEPHANIE ECKMAN, Institute for Employment Research (IAB), Nuremberg, Germany
(co-authored with Frauke Kreuter)

Efforts to reduce undercoverage of housing units, household member, and members of targeted subpopulations are often quite costly. Ideally efforts to increase coverage reduce undercoverage bias, but it is also possible that the newly covered units will be disproportionate nonresponders to the survey request. Several studies have demonstrated a possible connection between coverage and nonresponse. (For example, Peter Hainer's work with household roster coverage in the CPS, Tourangeau et. al.'s study of eligibility in screener surveys). We systematically evaluate this evidence and search for clues to the mechanisms underlying this connection.

At the conference, we hope to foster a discussion about the cost and quality aspects of the trade-off. While low-coverage high-response designs are probably the least expensive, the high reported response rates in such surveys serve to hide the low coverage rate. A high-coverage low-response design is almost certainly more expensive, but the reported response rate better captures the overall representativity of the survey. We will solicit feedback on what kind of experimental design would best allow us to test hypotheses about the connection between these two error sources.

Error Tradeoffs in Cross-National and Comparative Surveys: Questionnaire Design from a Total Survey Quality Perspective

GORDON WILLIS, National Cancer Institute, National Institutes of Health, United States

Several recent conferences and Workshops have addressed the issue of achieving cross-cultural equivalence of survey questions – especially when questionnaires are translated into different languages. It seems evident that efforts to control variation in the survey response process across cultural and language groups are of vital importance, as survey researchers are justifiably concerned that comparative analyses represent true effects, rather than persistent forms of measurement error across populations. However, based on experience with comparative studies, and on reasoning through the implications of the developmental work done in their support, I argue that we must take care to avoid potential adverse effects of these efforts, especially with respect to appropriate balance from a Total Survey Error (or a Total Survey *Quality*) perspective. Most broadly, the pursuit of cross-cultural comparability could end up sacrificing elements of the design that may influence either other components of variance or bias, timeliness, or overall "fitness for use" of the data. Sorting though these various effects, from an inclusive point of view that focuses on multiple error sources, is challenging, yet is not often not well-described or directly considered. In this paper I attempt to outline some major issues, with pointed examples.

My first consideration is based on the observation that designers of any complex system tend to believe that their part of the whole is the most important and should receive the lion's share of attention and resources. In the area of questionnaire design, the focus is often put on up-front determination that survey questions are translated properly, and that the questions are adapted culturally to each context. As such, some studies (e.g., Willis, et al. 2012, presented at the Comparative Survey Design and Implementation Workshop, Washington, DC) have put considerable resources into such development. This is laudable, but raises several questions concerning whether we pay sufficient attention to each of the following:

a) The degree to which we attend to the varied ways that questionnaires are *administered* across organizations and countries, in a way that may affect response rate and response quality;
b) Differences in *data coding* (e.g., coding of open-ended responses) that may produce systematic variation or bias across sites;
c) Variation in processes used to *assess data quality*, rather than to prevent error at the outset (e.g., the measurer-reducer distinction made frequently by Bob Groves).

In addition, I will discuss the issue of questionnaire design modifications as these may affect the utility, or "fitness for use" of the resulting data. It has long been observed that tradeoffs that ease the survey response task for respondents (e.g., reducing the number of response categories) can reduce the value of the resulting data for statisticians, given the loss in number of analyzable distinctions. These effects can become exacerbated in comparative studies. For example, it has been argued that because of the difficulty of creating correspondence across languages between graded categories such as 'never, a little of the time, some of the time, most of the time, and always', it would be favorable to instead reduce the response task to a simple Yes/No formulation. Although this solution may be attractive from the point of view of maximizing correspondence of cross-cultural interpretation (as Yes and No are likely to present comparable meanings no matter the target language), this solution would severely limit the nature of the analysis that could be carried out.

As a further example, Willis et al. (2012) determined that, in response to a question on experience with racial/ethnic discrimination, the item "*You worked harder to prove them wrong*" did not function well in languages other than English, as the translations to Spanish and Asian languages implied "*working at a job or business*," which was not the intent. In order to achieve *decentering* in such a way that all languages were coordinated, we changed the question in all languages to "*You tried harder to prove them wrong*." Although this change to the English version did reduce variation between languages, it created a substantive (though subtle) modification to the original English version. Hence, the effect this may have on the previously assessed psychometric properties of this item, as it contributes to an overall scale, is unknown. In any event, it becomes impossible to state, post-analysis, that the scale used was identical to the source instrument, prior to translation and adaptation.

These examples are not meant to imply that the decisions made were in error – or that their net effect was to increase overall survey error. Rather, the point is that decisions need to be made in the context of the overall effects they may have on data quality, and on fitness for use. I therefore propose that researchers attend to these issues, and will rely on the Conference as a forum for presenting this concept as an important area that is relevant to total survey quality of cross-cultural and comparative surveys.

A Tool for Managing Product Quality

PAUL BIEMER, Research Triangle Institute (RTI), United States
(co-authored with D. Trewin, H. Bergdahl, L. Japec, and Å. Pettersson)

In 2011 Statistics Sweden was presented with a challenge by its main stakeholder – the Ministry of Finance - to develop indicators that could show developments in product quality. There are numerous quality frameworks that address different dimensions, such as organizational, process and product quality. The challenge, however, is to measure and monitor changes in product quality in a comprehensive and systematic way and to clearly and concisely present progress on total survey quality to stakeholders. During three months a tool for this task was developed and tested on eight key products including the National Accounts and the Labor Force Survey.

Here we describe this tool and how it can be used to set clear measureable goals for product quality. Also, some results from the product evaluation and the associated recommendations for quality improvement are presented. Finally we summarize lessons learned and provide suggestions for modifying our approach.

Framework for Empirical Cost Modeling Relating Cost and Data Quality

MARY H. MULRY, U.S. Census Bureau, United States
(co-authored with Bruce D. Spencer)

At all times, but especially in times of increased cost constraints, it is important that statistical surveys and statistical programs more generally be cost-effective. A *cost model* is a tool for specifying the data quality (DQ) profile that is attainable for a given level of cost (or resources more generally); see in particular chapter 2 of R. M. Groves (1989) *Survey Errors and Survey Costs*. New York: Wiley. For successful practical application, cost models must be empirically based rather than merely theoretically constructed. Cost models are most useful when they are viewed as functions mapping costs to DQ profiles, because then they can be used in cost-benefit analyses relating costs to the benefits arising from data uses that are sensitive to DQ. The benefits can be quantified via loss functions.

DQ is inherently multidimensional, in the sense that DQ includes bias, variance, and timeliness even for a single statistic for a single subgroup at a single point in time. The dimensionality increases when statistics for multiple domains are considered, including subnational geographic areas such as states, counties, cities, etc. and subnational demographic groups, including not only racial or ethnic groups but subgroups used as controls for ratio-adjustment or other calibration in government surveys.

Considering data quality DQ (or its complement, total survey error (TSE)) to be multidimensional, we seek to attain a profile of DQ that is *Pareto-optimal*, i.e., no component of DQ can be improved without either increasing the resources available to the statistical program or decreasing another component of DQ. We may characterize a statistical program as cost-effective if the DQ is Pareto-optimal for the given resources. Resources may be expressed in terms of dollars, or person-hours, or time constraints, or combinations thereof. Therefore, cost is multidimensional as well. For a given cost, there may a set of Pareto-optimal profiles of DQ, and we will refer to that set as the DQ frontier available for the given cost. Thus, the general cost-modeling question becomes one of estimating the DQ frontier associated with any feasible cost vector (allocation of resources), along with, of course, the uncertainty in the estimated frontier.

Development of cost models must recognize that different data improvement efforts need not operate independently. Therefore, the empirical data underlying the cost modeling needs to be obtained from studies (whether observational or experimental) in which the various improvement efforts are varied simultaneously. For example, running two independent experiments, one comparing data improvement program A with a control and another comparing data improvement program B with a control, may be inadequate for predicting the DQ attainable when both programs A and B are implemented.

Previous work by Mulry and Spencer on the quality of small area estimation of population size in census years may be viewed as an empirical evaluation of DQ attained for a single set of resources. This paper expands on previous work to develop a framework for developing an empirical model relating DQ and cost when both are allowed to vary.

The World's Simplest Survey Microsimulator (WSSM):
A Tool for Understanding Total Survey Error

ALAN F. KARR, National Institute of Statistical Sciences, United States

In this paper we introduce the World's Simplest Survey Microsimulator (WSSM), a modular, extensible set of (C-language) computer programs whose goal is to move survey science in the direction of becoming a laboratory science. We then show how the WSSM can be applied in order to understand and reason about total survey error (TSE).

Using fewer than 5000 lines of source code, the WSSM generates a household-based population, including both frame and survey variables; simulates interviewers with varying skills and costs; selects a sample; conducts the survey, with, optionally, successive contacts via the Web, CATI and CAPI; incorporates both unit and item nonresponse; implements (a limited set of) edit rules and (a limited number of) imputation methods; computes inference-based data quality measures (Hellinger distances and Kullback-Liebler divergences) that relate survey estimates to the population "ground truth;" and calculates a variety of costs.

Essentially every component of TSE is present in WSSM--some more explicitly than others, and all can be varied. We illustrate how WSSM can be used to elucidate the contributions of, and relationships among, these components. We also show how WSSM can be used to support principled cost-data quality tradeoffs.

Disentangling mode-specific selection and measurement bias in social surveys

JAN VAN DER LAAN, Statistics Netherlands
(co-authored with Bart Buelens, Barry Schouten and Jan van den Brakel)

In 2011, Statistics Netherlands conducted a large-scale mixed-mode experiment linked to the Crime Victimization Survey (CVS). The objective of the experiment is to estimate total mode effects and the corresponding mode effect components arising from undercoverage, nonresponse and measurement for a number of key statistics from the Labour Force Survey (LFS) and the CVS. The estimated mode effects are used to improve methodology (data-collection strategies, nonresponse adjustment methods and questionnaire design) for mixed-mode surveys.

The experiment consisted of two waves: one wave with random assignment to one of the modes web, paper, telephone or face-to-face, and one follow-up wave to the full sample with interviewer modes only. The questionnaire is the CVS in which the last two blocks of questions are replaced by the LFS question block on employment status and a question block taken from the European Social Survey (ESS). Wave 2 is a follow-up of the full sample, excluding administrative nonresponse and nonresponse due to language problems and physical or mental illness.

In this paper, we define mode-specific selection and measurement bias, and we introduce and discuss estimators for these bias terms based on the experimental design. Importantly, these bias terms are estimated relative to a reference mode, in this case face-to-face. The proposed estimators are applied to key survey variables from the LFS and the CVS.

The main findings for the CVS are:

- Measurement effects dominate the mode effects for almost all target variables; coverage effects are mostly negligible.
- Web answers tend to be most negative (more victimization, less satisfied with police), while telephone gives the most positive picture.

The main findings for the LFS are:

- Employment and unemployment estimates for telephone and web differ significantly from those in face-to-face.
- There is no dominant mode effect component: coverage, nonresponse and measurement effects all play a role. The exception is web coverage and nonresponse effect for unemployment: web attracts significantly more employed respondents.
- Relative selection effects can be explained by standard variables from administrative sources, so that weighting should remove most of the mode-specific coverage and nonresponse.
- Mode effects for educational level indicate that some categories are difficult to classify by respondents.

Since measurement effects cannot be removed using weighting, also mode effects cannot completely be removed using weighting in the CVS. LFS mode effects can be explained by current weighting variables, i.e. the weighting model should be effective in removing mode-specific selection bias.

Evaluating Mode Effects in Mixed-Mode Data Using Back-Door, Front-Door, and Instrumental Variables

JORRE VANNIEUWENHUYZE, Catholic University Leuven, Belgium

Mixed-mode surveys are surveys where population members are allocated to different data-collection modes like CAPI, CATI, mail SAQ's and Web SAQ's. Because data-collection modes go with mode-specific selection error (the error introduced by researching a small subset rather than the entire population), it is argued that combining data-collection modes within one study increases the external validity and reliability of the obtained sample compared to samples obtained by single-mode surveys.

However, it is not guaranteed that data-quality increases when a mixed-mode design is used instead of a single-mode design. Indeed, the advantage of selection effects between the modes may be counteracted by differences in measurement effects between the modes (i.e. a difference in measurement error). Knowledge of selection effects and measurement effects helps evaluating mixed-mode data. Unfortunately, selection effects and measurement effects are completely confounded and estimation thus requires profound analysis.

This presentation aims to discuss three possible strategies to tackle the problem of confounded selection effects and measurement effects starting from the Causal Inference framework. These methods make use of back-door, front-door, and instrumental variables. Back-door variables are widely applied to mixed-mode data-analysis. Front-door variables and instrumental variables, in contrast, remain relatively unexplored within the analysis of mixed-mode data. Special attention will be paid to the assumptions of all three strategies and it will be argued that commonly used strategies fail to estimate the mode effects because of unrealistic assumptions. An example with mixed-mode data from a survey on opinions about surveys will be used for illustration."

Systematic and Random Error in a mixed mode Online-Telephone Survey:

An MTMM approach

EDITH DE LEEUW, Utrecht University, the Netherlands
(co-authored with Joop Hox and Annette Scherpenzeel)

To reduce nonresponse and coverage error at affordable costs, mixed-mode surveys are often advocated (e.g., De Leeuw, 2005). The final goal of mixed modes is to combine data from different sources, which assumes that data can be validly combined.
In the past, extensive mode comparisons have been made for the traditional data collection methods: face-to-face interviews, telephone surveys, and self-administered paper mail questionnaires, suggesting a dichotomy of survey modes in those with and without an interviewer. There are far fewer comparisons of Internet with interview surveys (cf De Leeuw & Hox, 2011).

A main problem in mode comparisons is the criterion on which modes are compared as accuracy of measurement is operationalised in different ways in different fields (cf. Groves, 1987). Ideally, the true value is known and this true value is then compared with the reported value. Some studies do have access to a validation criterion and therefore can carry out a record check (e.g., Krauter, Presser & Tourangeau, 2008). But, when subjective phenomena are studied, hard validation data are per definition not available, and researchers have to rely on various proxy indicators of data quality. Biemer (2001) points out that these often rely on assumptions on the direction of the biases (e.g., overreporting of sensitive information) and argues in favour of a model-based approach instead. A direct model-based approach to the analysis of measurement error in surveys on subjective phenomena is the multitrait-multimethod (MTMM) design that allows separation of substantive or trait variance, method variance, and error variance (Campbell and Fiske, 1959; Alwin, 1974; Saris and Andrews, 1991).

We used the MTMM-approach in a mode-comparison implemented in the Dutch LISS-panel, which is a high quality, probability-based Internet panel. Panel members were randomly assigned to one of two modes: a computer assisted telephone interview or a web survey. Mode of data collection is the method factor and within each mode the same five questions (traits) were posed. One month later the same respondents were asked the same questions, but now in a uni-mode (web) survey. This design enables us to disentangle systematic and random error in both telephone and web survey and to investigate if measurement equivalence holds over modes.

Inference in Surveys with Sequential Mixed-Mode Data Collection

JAN VAN DEN BRAKEL, Statistics Netherlands
(co-authored with Bart Buelens)

The use of mixed data collection modes receives increasing attention by national statistical institutes. Driving factors behind the rising importance of mixed-mode designs in survey sampling are the increasing pressure to reduce administration costs, attempts to reduce non-sampling errors and new technological developments which leads to new data collection procedures. Sequential mixed-mode strategies applied at Statistics Netherlands typically starts with Web Interviewing (WI) and uses interviewer-administered modes in the follow-ups to contact the remaining nonrespondents. WI combines the benefits of traditional self-administered data collection modes through paper and pencil with the power of computer assisted administration, i.e. cost effective, greater sense of privacy for the respondent and the possibility of a more complex questionnaire design. The follow-ups are based on computer assisted telephone interviewing (CATI) and computer assisted personal interviewing (CAPI) and is required to guarantee sufficient coverage and response rates.

Mixed-mode surveys are known to be susceptible to selection effects and mode-dependent measurement errors, collectively referred to as mode effects. While the use of different data collection modes within the same survey may reduce selectivity of the overall response, it is characterized by measurement errors differing across modes. Inference in sample surveys generally proceeds by correcting for selectivity -- for example by applying generalized regression -- and ignoring measurement error. When a survey is conducted repeatedly, such inferences are valid only if the measurement error remains constant between surveys. In sequential mixed-mode surveys, where non-respondents in one mode are re-approached using a different mode, it is likely that the mode composition of the overall response differs between subsequent editions of the survey. Variations in the mode composition lead to variations in the total measurement error, invalidating classical inferences. An approach to inference in these circumstances is proposed in this paper. First, it must be ascertained that the response is appropriately corrected for selectivity. Second, the mode composition of the response is calibrated towards fixed levels. Assumptions and risks associated with such a procedure are discussed. An example from the Dutch Crime Survey, based on WI, PAPI, CAPI and CATI is used throughout the paper to illustrate the proposed approach.

Evaluating the Extent of Non-Response and Non-Coverage Bias in the Swiss European Social Survey

CAROLINE ROBERTS, University of Lausanne, Switzerland
(co-authored with Caroline Vandenplas and Michèle Ernst Stähli)

Surveys increasingly require expensive response enhancement methods to achieve target response rates. In the European Social Survey, the specifications for participating countries stipulate a response rate target of 70%, intended originally as a way to improve cross-national equivalence, guide sample design, as well as to encourage the pursuit of high rates of participation, generally accepted as the best way to minimize bias from nonresponse. Yet the specification of such a target on the ESS has long been the focus of controversy, mainly due to mounting evidence that the relation between response rates and nonresponse bias is not as clear cut as it was originally assumed to be, and that the efforts needed to increase participation – as well as becoming increasingly unaffordable – may in fact aggravate the problem of bias on certain survey variables, and threaten comparability across countries. This literature has not only called into question the value of setting response rate targets, but has highlighted the need for alternative indicators of survey quality that take into account the extent to

which data are affected by different forms of survey error. One of the proposed measures is the 'R-indicator' or Representativeness-indicator (Schouten and Cobben, 2007[1]), which relies on estimates of individual response propensities based on available auxiliary data, to assess the representativeness of the responding sample. In this paper, we draw on this literature to assess the effectiveness of different response enhancement strategies used on the European Social Survey in Switzerland, both in terms of costs, as well as in terms of their impact on errors due to non-response and non-coverage.

After a comparatively low response rate was obtained in the first wave of the ESS, significant improvements have been made in Switzerland, but at considerable financial cost. Until recently, insufficient data relating to the non-responding sample were available to enable a detailed cost-benefit analysis of the effectiveness of methods used to increase participation. In the most recent wave of the survey, however, the availability of new data means that such an analysis is now possible and desirable. Since 2010, the Swiss Centre of Expertise in the Social Sciences (FORS) has been allowed (under strict conditions) to draw their survey samples from the Federal Office of Statistics population register, which covers all (registered) people resident in Switzerland. This new frame offers several distinct advantages: it decreases error due to non-coverage substantially compared to past social surveys, which generally relied on telephone lists for sampling purposes; it also permits the sampling of individuals, thereby reducing the likelihood of refusals by proxy; and importantly, it means that for the first time, auxiliary data are available for the analysis and possible correction of coverage and nonresponse error. The Swiss ESS 2010 was one of the first studies to profit from this new sampling frame, and is somewhat unusual in terms of the richness of the data available for this kind of analysis. For each sampled person, we possess information about basic socio-demographic variables like age, gender, marital status, nationality, and address, and we also know if this person has a registered phone number or not. Additionally, paradata are available relating to fieldwork effort (the number, timing and outcome of contact attempts), and interviewer observations (e.g. about the condition of housing, neighbourhood characteristics and an evaluation of wealth). Finally, about 50% of the non-respondents accepted to fill a non-response follow-up paper questionnaire that was sent to them a few months after the end of the main data collection, providing information about the extent to which key survey variables (including attitudinal measures) may be affected by bias.

To evaluate the effectiveness of different fieldwork strategies used on the ESS 2010, we use R-indicators to measure the representativeness of different sub-groups of respondents, participating after different amounts of fieldwork effort (based on the number of contacts needed to obtain participation, the need for refusal conversions, etc.). We also assess the representativeness of the participants to the non-response follow-up survey. Moreover, we are interested in the cross-section of these subgroups and the subgroups of sampled persons having a registered telephone number. This information is interesting for at least two reasons. First, it may provide insight into the extent of coverage error in previous survey waves, when samples were drawn from the Swiss telephone-register, which is known, to suffer increasingly from under-coverage. Second, telephone contacts are permitted for sample members with available telephone numbers after five unsuccessful face-to-face contact attempts. This difference in contact mode between people with and without a registered phone number could also introduce bias that could be detected using R-indicators. Finally, partial R-indicators will be calculated for different subsets of participants (e.g. after five face-to-face contact attempts) to distinguish under-represented groups in different steps of the fieldwork process.

[1] Schouten, B., Cobben, F. (2007), R-indexes for the comparison of different fieldwork strategies and data collection modes, Discussion paper 07002, CBS, Voorburg.

Preliminary analyses suggest that the 'representativeness' of the respondents compared to the sample, based on the auxiliary variables available in the frame (e.g., gender, age and urbanization) does not necessarily improve with the number of contacts attempts and hence with increasing response rates. On the other hand, including the respondents to the non-response survey significantly increases the value of the R-indicator. These first results will be extended to other auxiliary variables available, for instance, from the interviewer observations, and crossed with the subset of people in the sample who have a registered number. We will also compare the final group of respondents to the main survey with the set that includes participants to the non-response survey, to allow us to compare R-indicators based on socio-demographic auxiliary variables with indicators incorporating attitudinal variables collected in the non-response survey. Ultimately, our aim is to gain insight into how to develop targeted contact or refusal conversion strategies for different types of persons based on knowledge of the representativeness of the responding sample, with a view to reducing costs and survey error. Such a targeted fieldwork strategy could help to improve data quality compared to blind efforts to increase the response rate to meet internationally required thresholds.

On the Relationship Between Non-Response Error and Measurement Error in Response Enhancement. The Norwegian Election Survey System as a Case Study

ØYVIN KLEVEN, Statistics Norway
(co-authored with Ib Thomsen and Li-Chun Zhang)

It is a well established practice in most NSIs that chasing reluctant or hard to get respondents can reduce sampling variance and bias caused by non response. We have previously demonstrated that "hard to get respondents" in the Election Survey differ from the easy to get respondents on several key estimates from the survey. By using administrative data from the local electoral offices we have a long history of controlling the non response bias on the electoral turnout. The relationship between measurement error and response enhancement like chasing non respondents has to our knowledge not been analyzed to the same degree. Our concern is that when we are chasing non respondents we might increase other sources of error like measurement error. By linking data on each sampled unit from the survey to in our Election Survey System we have the possibility to throw some light on this topic. We can use sample-linked administrative registers like voted/not voted level of education and income to compare answers given on the same topic in the survey. Because the survey is a rolling panel we can also compare answers given to the same topic by the respondents at different waves. Paradata who identify several characteristics of the respondents and non respondents, like easy to get hard to get, is also linked to the data file. We will use the Election Surveys from 1997, 1999, 2001, 2003, 2005, 2007 and 2009 as our main empirical base but we will also include other surveys.

Examining Interviewer Behavior in Handling Difficult Cases

WENDY HICKS, Westat, United States
(co-authored with Aaron Maitland)

Interviewers play an important role in gaining the cooperation of survey respondents and administering questions. Several studies have explored the relationship between interviewer behavior and different sources of survey error. Olson and Brady (2010) decompose interviewer variance into measurement error and nonresponse error. They find that interviewer related variance can be due to both differences in the characteristics of respondents across interviewers and also measurement difficulties. Tourangeau, Kreuter and Eckman (in press) demonstrate interviewers' effect on both nonresponse and coverage error. But there is little known about the mechanism in which interviewers' affect error.

A study by Olson and Peytchev (2007) adds some insight into the interviewers' effect. The authors found that as a study progresses, and interviewers conduct more interviews, the length of the interview decreases and the interviewers perceive the respondents as less interested. While we would anticipate that interviewers improve their skill in navigating and administering an instrument over repeated administrations, the change in interviewers' perception of respondents' interest in the study may not be independent of the faster administration and may actually be more reflective of the interviewers' own attitude. In addition, the authors control for respondent differences and conclude that respondents later in the field period do not differ significantly from those who participated earlier in the field period in regard to interest in the survey topic.
In this paper, we build on these findings and make use of Computer Audio Recorded Interviewing (CARI) and coding analysis to further understand the mechanism in which interviewers' behaviors may affect error. Surprisingly, previous studies using behavior coding analysis have not found a strong or consistent relationship between interviewer behavior and data quality (Hess, Singer and Bushery, 1999; Dykema, Lepkowski and Blixt, 1997). This is counter to conventional survey practices in which we spend non-trivial time and money training interviewers, monitoring them and coaching them in order to increase their adherence to study protocols, all with the assumption that without this emphasis, interviewers will in fact negatively affect data quality in terms of nonresponse, measurement error or both.

Schaeffer and Dykema (2011), offer a brief review of studies that have tried to link interviewer behavior and data quality, and suggest that different coding variables, or different models are needed to properly identify the relationship between interviewer behaviors and survey error. Using the National Health and Aging Trends Study (NHATS), we link behavior coding analysis with contact history data and interviewer characteristics to create more context in which we examine the relationship between interviewer behavior and data quality. In the analysis, we construct a 'case difficulty' variable based on the contact history data and compare interviewer behaviors between the more difficult and less difficult cases. In addition, we account for interviewer productivity as a variable related to interviewer behaviors. In a preliminary analysis, we found that interviewers differ in how well they follow the standardized interviewing protocol between difficult and less difficult cases, depending on their overall productivity. Less productive interviewers tend to deviate from the standard interview administration protocol more often with more difficult cases (X^2=5.27, p<0.05). This is in fact the direction we expect and in fact may encourage with some interviewer strategies implemented at different points in the field period. For example, as we move into the final stages of data collection which often includes a refusal conversion effort, survey organizations may allow for reduced protocols or other short cuts in order to persuade the more reluctant respondents to participate. It is conceivable that this emphasis on an 'easier' or shorter interview to facilitate cooperation may blur the lines for interviewers in terms of acceptable interviewing behaviors.

However, somewhat surprisingly, interviewers at higher levels of productivity demonstrate greater deviations from the expected protocol for the less difficult cases (X^2=8.67, p<0.05). And this relationship exists even when accounting for the order of interviews conducted across the field period. The more productive interviewers demonstrate less standardized interviewing behaviors more often with less difficult cases, consistently during the field period.

In this paper, we continue to explore the CARI data to gain further insight into why and how interviewers deviate from protocol, by case difficulty. In addition, we look at whether there are differences in data quality as measured by item nonresponse, interview length and the consistency of survey responses when interviewers' deviate from protocol, controlling for case difficulty.

What is the Impact of Mode Effect on Non-Response Survey Usability?

CAROLINE VANDENPLAS, University of Lausanne, Switzerland
co-authored with Dominique Joye and Michèle Ernst Stähli

In many countries across the world, designers of social surveys have to face a growing problem of non-response and increasing costs to reach and convince people to participate in their study. The problem with non-respondents, including non-contacts and refusals, is that they do not always represent a random sub-group of the sampled persons; their characteristics, e.g. socio-demographic, socio-economic or, even worse, behavioural and attitudinal, differ from the characteristics of the participants. This implies that the results inferred from the respondents may potentially be biased. Sadly, we often have no or little information about the non-respondents. Moreover, even though it is expected that higher response rates lower the risk of bias, this relation is far from being linear (Groves 2006; Groves and Peytcheva 2008). The bias is therefore difficult to predict and estimate. Further, the closer the variables that determine the propensity to participate are to the actual subject of the survey the higher the bias (Billiet et al. 2009). Researchers have therefore increased their effort to collect information about the non-respondents to try to find out what are the motivations behind refusals and the reasons for non-contacts. Non-response surveys could be a solution as they can be conducted at relatively low costs (self-administered questionnaires). The purpose of this paper is to assess if this low cost mode does not compromise the usability of the collected data, e.g. mode and time effects.

One way to deal with non-response bias is to calculate non-response adjustment weights based on post-stratification. The variables commonly used to construct post-strata are socio-demographic variables, mainly because they are generally the only available ones. Demographic variables are however known to not be the best predictors of response patterns and hence, one can be sceptical about the degree to which this weighting procedure really corrects for non-response. It should be clear that the more related the variables are to the probability to answer, the better they are to adjust for non-response bias. For instance, social involvement and political interest are often more correlated with participation outcomes (Matsuo et al. 2010) than age or gender. The difficulty is to find the balance between the possibility of collecting relevant information (e.g. attitudinal) about non-respondents, the costs of doing so, and the potential gain in corrections for non-response bias.

In Switzerland, non-response surveys have been conducted for several different social surveys: European Social Survey (2006 and 2010), European Values Study (2008) and the MOSAiCH 2011 ('Measurement and Observation of Social Attitudes in Switzerland' comprising the Swiss version of the International Social Survey Programme (ISSP)). The purpose of those studies is to better understand the profile of non-respondents, not only based on socio-demographical variables but also on attitudinal characteristics. In every case, the non-response survey was designed as a short self-administered paper questionnaire (max. 15 questions) which was sent by post in a time frame which varied from two to six months after the concerned individuals were initially contacted for the main data collection (all the non-respondents qualified to be part of the non-response survey as well as a control group of 300 respondents). Overall, the response rate varied between 55 and 65%. This questionnaire repeated some of the items from the main study and collected information on the possession of a telephone number and composition of the household. The aim was to use this information to better understand non-response bias in the main surveys and to calculate propensity scores to use in non-response adjustment weights (Matsuo et al. 2010).

The variables that explain the best the propensity to answer have proven to be stable across cycles and surveys (Joye et al. 2011). In general, they relate to themes of 'social involvement', 'political trust' and 'attitude towards immigration'. There remain the issues of the different modes in which the main and corresponding non-response surveys have been conducted and the time frame shift (Beullens et al., 2009). Indeed, in a few months, public and personal events can change the opinion, and hence, the answer given by a person to a specific question. Moreover, on top of possible seasonal effects, completing a self-administered paper questionnaire or facing an interviewer can influence the given answer, especially on sensitive questions concerning politics or immigration. To control for these effects, the non-response survey has also been administered to a selected group of respondents.

The purpose of this research is to, first, study in how far this compromises the utility of the non-response survey. We will, in a first step, investigate the possibility to correct for the mode and time difference effects and identify the variables that can be used for adjusting any non-response bias. In a second step, we will calculate propensity scores to adjust for non-response bias, based on the variables that are stable across studies but also across the main survey and the corresponding non-response-survey, eventually controlled for mode and time effects. We will then verify our results by comparing the results of key variables common to both (main and non-response) studies before and after weighting and checking whether significant differences between participants and non-participants have, as expected, disappeared.

A Multi-Method Analysis of the Relationship between Item Refusal and Measurement Error, Using a Measure of the Public's Trust of Official Statistics in the United States

MORGAN EARP, Bureau of Labor Statistics, United States
(co-authored with Jennifer Hunter Childs, Melissa Mitchell and Stephanie Willson)

In an effort to explore the public's trust of official statistics in the United States and attitudes towards the use of administrative records, the Census Bureau collaborated with several federal statistical agencies to develop a measure of trust in statistical products, trust in statistical agencies, and attitudes towards use of administrative records. This measure is being used to monitor the public's trust level and assess the impact on attitudes towards use of administrative records.

During the construct and item development phase, we consulted international models of trust of official statistics (Brackfield, 2011; UK Office for National Statistics, 2006 & 2007). Prior to data collection (and during data collection), cognitive interviews and expert reviews were used to assess and improve items. Pilot data was collected in three phases, allowing us time to assess and address measurement error between administrations. During all three pilot phases of data collection, we used random probes to assess item performance and confirmatory factor analysis (CFA) to evaluate item misfit (error variance) within factors. Using a combination of cognitive interviews, expert reviews, random probes, and CFA, we detected items suffering from measurement error and made recommendations for improving and/or removing items.

This paper focuses on the relationship between the various diagnostic tools used to assess measurement error and the relationship between measurement error and item nonresponse. We will present the theoretical model we developed, the methods used to detect measurement error, and our analysis of the relationship between item nonresponse and measurement error.

Combining Predictive Modeling and Operational Insights for Effective Online and Face-to-Face Recruitment in Urban and Rural China

YU-CHIEH LIN, ISR - University of Michigan, United States
(co-authored with Teresa (Ye) Jin, Shu Duan, and Jennie W. Lai)

Household panel recruitment can be challenging in the emerging markets where households in the underdeveloped or rural areas may have limited understanding on the concept of survey research. In consideration of growing Internet usage in China, online recruitment method can be leveraged for particular hard-to-reach segments in China. In consideration that China is a collection of diverse regions with economic growth, technology development, cultures, dialects, consumer behaviors, and lifestyles, strategies of offline and online recruitment used vary in different tier of cities. The Nielsen Company has deployed a mixed-mode method of recruiting households online as well as face-to-face for a consumer panel to study their purchasing behavior. This research paper will focus on the advantages and disadvantages of the mixed-mode recruitment approach by analyzing both quantitative and qualitative data collected from each mode. The second part of this research paper will focus on panel attrition and maintenance among recruited households. The maintenance strategies, participation history, duration in the panel, and activity status of households recruited by mixed-mode will be reviewed and discussed. Multivariate modeling technique will be used to potentially predict the likelihood of households to cooperate and stay in the panel using data collected online and face-to-face recruitment such as paradata, household/respondent demographics, purchasing behavior, lifestyle variables, geography, etc. Furthermore, the para data will also be examined to evaluate the efficiency of recruitment method by mode. Finally, qualitative interviews will be conducted with the field recruiters to gather insights on operational challenges and best practices of the recruitment process and tools provided.
These key research findings and recommended best practices will be shared in hopes of shedding light on an effective and efficient method of recruiting hard-to-reach segments in mainland China.

Challenges of Assessing the Quality of a Prerecruited Probability-Based Panel of Internet Users in Germany

BELLA STRUMINSKAYA, GESIS - Leibniz Institute for the Social Sciences, Germany
(co-authored with Lars Kaczmirek)

To answer methodological questions about the optimal recruitment and maintenance of probability-based online panels, GESIS conducted a pilot project, in the course of which Internet users were recruited by telephone to participate in an online panel. The recruitment was based on a dual frame RDD sample, which included both landline and cell phone numbers. After a short telephone interview, respondents were asked to provide their email address in order to participate in an online panel. Willing respondents were to complete an online questionnaire of about 10 to 15 minutes duration every month for the total period of eight months.

One of the goals of the GESIS Online Panel Pilot project is to evaluate the quality of data collected online and the overall quality of the panel. One way of doing so is to compare the survey estimates to official records or other external sources of information, where one particular statistic is present (i.e. Scherpenzeel and Bethlehem 2011, Yeager et al. 2011). However, this method is problematic when official sources do not contain information on a target population. This is the case with the GESIS Online Panel having a target population of German-speaking Internet users over 18 years of age since the share and the characteristics of the Internet using population in Germany are themselves subject to estimation (Destatis 2011).

In the absence of such "gold standard" benchmarks, estimates from other general population surveys offer an alternative course of quality assessment. In our case these surveys are the German General Social Survey (ALLBUS) and the German part of the European Social Survey (ESS), which both include information on private Internet usage and had fieldwork carried out within a similar timeframe as the fieldwork of the Online Panel. The reasons for treating these surveys as benchmarks in terms of data quality are 1) higher response rates and, more importantly, 2) selective nonresponse in the multistep recruitment of the online panel, showing demographic and attitudinal differences between those willing and unwilling to participate in the panel as well as differences between those starting the online questionnaire and those not responding to the online survey.

A straightforward procedure of comparing estimates from the Online Panel to the reference surveys and applying *t*-tests for statistical significance of the pairwise comparisons provides mixed results. Some of the differences are consistent across surveys: significant differences between respondents of the online panel and reference surveys in age (younger in the Online Panel), no differences in gender composition, higher levels of education in the online panel and significantly more singles than in both of the other surveys. Some of the other differences are inconsistent: more respondents working for pay are found in the Online Panel than in ALLBUS but no such differences are found in comparison with ESS. Fewer individuals with immigration background are found in the Online Panel compared to ALLBUS, but not compared to ESS. Differences are also found in some attitudinal aspects: higher life satisfaction in ESS, different health status in ALLBUS (rating they are in very good health).

Some concepts are operationalized differently in ESS and ALLBUS. Here, we had to decide which survey had to be used as a reference. For such cases, either ESS or ALLBUS could be compared to the Online Panel. The results also seem to be inconsistent. Differences with ALLBUS are found in political interest, with ALLBUS having more respondents very interested in politics. Respondents in ALLBUS are also less trusting than Online Panel respondents. Online respondents rate the current German economy a little better than ALLBUS respondents, and next year's economy – worse. No differences are found in ratings of respondents' own economic situation for the current and next year. With respect to ESS no significant differences are found in satisfaction with government. However, online respondents are less satisfied with the health system, with their job, and with the balance between work and free time.

One important limitation to this approach of assessing the data quality is the choice of the set of variables to be examined. It seems that the procedure of comparing means and reporting absolute errors with respective significance tests is performed in the majority of studies which examine the quality of online panels. However, apart from demographic variables, covariates considered by researchers range broadly from voting behavior or attitudes to immigration to smoking behavior. The inherent problem of performing such multiple comparisons with various covariates is the concern about finding more differences when more variables are added to the set. It seems unlikely that a set of covariates may be found, where it could be agreed that the given variables reflect the data quality. In the practical setting the choice of variables for comparison seems to be dictated by the availability of external benchmarks. This raises a question about an appropriate method of analysis, which can produce comparable results across studies.

Questions for discussion:

1) Do participants of the workshop have ideas on how to restrict the set of measures for analysis or alternative (statistical) methods of analysis, which would be appropriate for the assessment of data quality?

2) How to approach the problem of the two benchmark surveys, similar in terms of mode, target population, and fieldwork period, producing inconsistent results when compared to the online panel?

Placement, Wording, and Interviewer Effects: Experiments in Obtaining Data Linkage Consent from Survey Respondents

VALERIE TUTZ, Ludwig Maximilians University of Munich, Germany
(co-authored with Joe Sakshaug and Frauke Kreuter)

Data linkage is getting more and more important. It’s a time and money saving method to get a lot of data with much information about one respondent. There are several types of data procedures. Some are statistical methods, what means that they use algorithms to find a “statistical twin”. The other method is an individual one, where the respondent’s data is exactly linked by some identification number. Because of data privacy issues, consent of the respondent is needed for individual linking. Not all respondents consent to linkage, which can lead to a biased data sample. To reduce the risk of bias, it is important to achieve high consent rates. This begs the question of how to increase the consent rate. Studies have found that respondent demographic characteristics have an effect on linkage consent, but those are factors we can't influence as a survey designer. An open area of research is to identify survey factors that we can influence while running a survey.

The purpose of this paper is to study survey design factors that can influence consent. The study of interest was conducted on a population-based sample in Germany from August to October 2011. Additional to other questions it contained a wording experiment and a placement experiment concerning the consent question. The wording experiment presented two different wordings both placed at the front of the questionnaire. The first wording in English said:
“The Institute for Employment Research in Nuremberg could merge the study results with data about your past times of employment, unemployment and participation in measures during unemployment. To connect this data with the data from the interview we would appreciate if you give us consent. Do you consent?”

The other wording had an addition at the beginning that says: *"To keep the interview as short as possible..."*. We hypothesized that those respondents who were given the “shorter interview” wording would be more likely to consent than the others. However, the expected result didn’t show up as the consent rates of both wordings were not statistically significant.

We also tested whether a placement effect exists. We observed that most of the studies using linked data place the consent question at the end of the questionnaire. We hypothesized that if placing the consent question at the front could improve the linkage consent rates. Perhaps respondents are in a "yes"-mood at the beginning. They just said "yes" to participate in the study and haven't given any information yet. So they may keep on saying "yes" again. At the end, it's possible that respondents are tired and think they already shared very much information, why should they share even more? To compare the consent rates of both placements, the same wording without the additional “shorter interview” phrasing was used.

The big difference can already be seen by comparing the consent rates. The consent rate for the front placement is 95.6% whereas the consent rate at the end amounts to 86.0%. This is a highly significant difference of more than 10%.

Another interesting factor of influence could be the interviewer. It is very interesting to see if the consent rates differ between interviewers and why. It is very probable that the interviewer motivation could affect the respondent's answer to linkage. If the interviewer is very friendly and tries to convince the respondent about the importance of the interview and the consent to the linkage question, the respondent could be more likely to do that. Also the interviewer's attitude could have an impact on their consent rate. If the interviewer herself would consent to linkage, she may explain the linkage consent theme more positively than otherwise. To address these issues, the interviewers were asked to complete a special questionnaire that included items asked about socio-demographic questions, questions about their personality, their motivation and about their attitude and expectation towards consent.

The consent rates for the individual interviewers varied from 75% to 100%. To explain this variation, we performed a logistic mixed regression model where the interviewers comprised random effects and the respondents' answer to linkage consent is the dependent variable. We found that interviewers with higher income have higher consent rates, interviewers who expect higher consent rates get a little bit lower consent rates than those who expect worse consent rates. Those interviewers who are more likely to consent themselves are also more likely to get better consent rates. More experienced interviewers get lower consent rates than inexperienced interviewers and men get higher consent rates than women. In conclusion, we observed that altering the standard wording of the consent request did not produce affect the consent rate. However, placing the consent question at the front of the questionnaire produced higher consent rates than placing them at the end, a design feature that could be easily implemented in practice. We also found interviewer factors, including experience, attitudes and expectations towards consent to be related to respondents' likelihood of consent.

Social Network Analysis as a Tool for Assessing Respondent Burden, Measurement Error and Nonresponse in Establishment Surveys

DIANE K. WILLIMACK, U.S. Census Bureau, United States
(co-authored with Alfred D. Tuttle)

Establishment survey respondents often must obtain assistance from other persons in their organizations in order to complete a survey request. This is especially true in larger organizations where the requested data may be distributed among multiple organizational units, information systems and/or personnel with specialized knowledge or system access. Qualitative research results suggest that finding and obtaining the assistance of other (secondary) respondents contributes significantly to the perceived burden of a survey request. Lorenc (2007) has also hypothesizes the prospects of associated measurement error. Moreover, research by Keller et al. (2011) suggests that respondent role is associated with data quality and item nonresponse. While many survey methodologists working on business surveys recognize this need for multiple respondents, there has been no systematic approach to measuring this phenomenon and its consequences for burden, measurement error and nonresponse.

In this paper we attempt to demonstrate a unique approach using social network analysis (SNA) methods to obtain quantitative and qualitative data on the primary and secondary downstream respondents who prepare an organization's survey response. In our approach we consider:

- The organizational role of the primary respondent;

- The number of secondary respondents who contribute to a given survey request and their roles in the company;
- The modes of communication used to communicate the survey request to them;
- The substance of this communication – i.e., the degree to which the content of the communication includes the actual survey questions, definitions, and instructions, or some prior interpretation of them.
- The secondary respondents' prior relationships, if any, to the primary respondent;
- The primary respondent's perceptions of the cooperativeness of and quality of data provided by secondary respondents;
- Whether primary respondents attempt to validate aggregate responses from multiple secondary respondents, for example by comparison to a known control figure;
- Whether all survey requests are channeled through a single individual or office, or whether they go directly to various respondents;
- The degree to which collection from secondary respondents is managed or monitored by a single office.

By gathering quantitative data on types and numbers of respondents contributing to responses, we expect SNA will aid in producing an objective indication of the burden imposed by a survey request. Such data may also provide indications of survey items that may not be performing well, contributing to measurement error and/or nonresponse, and thus merit further investigation.

More broadly, we can better understand the role organizational context plays in survey response, enabling us to adapt our survey questions, questionnaires, other respondent-centered response aids and communication strategies to natural record-keeping practices, information-sharing processes, and working relationships within businesses. We will demonstrate the utility of SNA methods using common scenarios observed in interviews with respondents at large companies receiving multiple surveys from the US Census Bureau. We will also discuss implications for measurement error, nonresponse and burden that may be associated with these behavioral networks among multiple respondents, and suggest possible strategies for improving the efficiency of survey implementation.

A Model-Based Procedure to Evaluate the Relative Effects of Different TSE Components on Structural Equation Model Parameter Estimates

DANIEL OBERSKI, Tilburg University, The Netherlands

The study of total survey error has mostly focused on univariate statistics such as means, totals, and proportions. Multivariate statistics, which can be often be more generally formulated as parameters of structural equation models (SEM), are often also of interest, however. For example, correlation, regression, differences between subclass means, and instrumental variables are all special cases of SEM.

The effects of individual survey error components such as measurement error or clustering on such models are well-known (e.g. Fuller 1987, Scott & Holt 1982), but to my knowledge no studies investigate the effect of different error sources on SEM parameter estimates simultaneously.

Such information would be useful to have, since it could give an indication as to the relative importance in terms of mean squared error of controlling different survey error sources. For example, if the purpose of the study is to estimate some parameters of a multivariate model and measurement error turns out to be highly influential relative to sampling error for that model, it could conceivably be cost-effective to allocate more of the budget to questionnaire development while sacrificing sample size.

I outline a simple model-based procedure for simultaneously estimating the relative efficiency of SEM parameter estimates due to different survey error components. The effects of clustering, measurement error, and non-normality are jointly estimated for an example multivariate model involving reciprocal effects, instrumental variables, correlated error terms, and measurement error on real data from the European Social Survey 2008.

It is shown how estimates of the effects of these different survey error components can be obtained under certain assumptions. In the example given, it is also shown that the relative sizes of these effects are very different than commonly found in the estimation of means and totals. In particular, measurement error is a much more important factor in our multivariate example than it would be for a mean or total. Some possible applications of the method are discussed, as well as its shortcomings.

Analytic Error as an Important Component of Total Survey Error: Results from a Meta-Analysis

JOE SAKSHAUG, Institute for Employment Research (IAB), Germany
(co-authored with Brady T. West)

The survey methodology literature is replete with alternative descriptions of the Total Survey Error (TSE) paradigm. One of the earliest descriptions of this paradigm can be found in the 1944 article entitled "On Errors in Surveys" by W. Edwards Deming in *American Sociological Review*. A special issue of *Public Opinion Quarterly* published in 2010 was dedicated to recent research on TSE, and includes two comprehensive overview articles (Biemer, 2010; Groves and Lyberg, 2010) presenting historical perspectives on these alternative descriptions. The majority of these descriptions essentially divide TSE up into four types of errors than can arise in surveys: coverage error, nonresponse error, measurement error, and processing error. While further divisions of these errors based on observation vs. non-observation and bias vs. variance are certainly possible, most of the published descriptions of TSE fail to recognize a very important source of error that is entirely out of the control of the survey researcher: analytic error, or a failure of the survey data user to employ appropriate estimation methods when analyzing the collected survey data. Recent publications have started to consider this aspect of TSE in greater detail. Groves et al. (2009) include analysis as a component of their "twin inferential process" description of total survey error. Biemer (2010) hints at analytic error as a form of processing error, and Smith (2011) prominently features analysis in a recent paper focused on refining the TSE perspective. These descriptions aside, the relative contribution of analytic error to TSE remains a gap in the collective knowledge of survey researchers.

Survey organizations often strive to minimize important sources of TSE (often at significant expense to funding agencies and the tax-paying public in general). However, these costly efforts will be for naught if users of the data fail to employ appropriate estimation methods that recognize features of the sample design that gave rise to the set of survey respondents. This problem becomes especially serious when secondary analysts of publicly available survey data submit articles presenting applied research for publication, and these analytic errors are missed by otherwise well-meaning reviewers in the peer-review process employed by reputable journals. As a result of this process, even the highest quality survey with all sources of TSE minimized could lead to publications that present error-prone population estimates.

With this study, we sought to quantify the prevalence of these types of analytic errors by performing a meta-analysis of the published literature from a variety of fields that perform secondary analyses of survey data arising from complex samples. As a secondary objective, we sought to explore whether characteristics of the journals in which these articles were published (e.g., impact factor, presence of statisticians on the editorial boards, analytic guidelines for authors, etc.) were related to the prevalence of various errors. Using online search tools (e.g., Google Scholar), we identified published articles from a variety of fields (e.g., public health, cardiology, mental health, sociology) presenting analyses of survey data collected from large, nationally representative samples in the United States (e.g., the National Health and Nutrition Examination Survey, or NHANES, and the General Social Survey, or GSS). With the assistance of graduate student research assistants, we then coded these articles on a variety of error indicators. Examples of these indicators included appropriate use of weights for estimation, appropriate use of sampling error codes for variance estimation, use of appropriate software for survey data analysis, appropriate subpopulation analysis approaches, and use of appropriate language to describe the results (e.g., weighted estimates vs. sample estimates).

While data collection is ongoing at the time of this submission, initial analysis results suggest that several types of analytic errors are quite prevalent, including inappropriate subpopulation analyses and a failure to use appropriate software. Analysts also fail to incorporate weights or compute standard errors reflecting sample design features more often than would be desirable, and we find that descriptions of analysis results and inferences may tend to mislead readers about the scope of the inferences (i.e., population vs. sample). We also find that most peer-reviewed journals, including those with large impact factors, fail to emphasize the use of specialized analysis methods for secondary analysts of complex sample survey data in their guidelines for authors. These initial results suggest that academic journals and survey organizations could do more work in emphasizing the use of appropriate analyses of a given survey data set. We would look forward to receiving feedback on this ongoing study and its potential contribution to the literature on TSE at ITSEW 2012.

A Total Survey Error Analysis of an Address-Based Sampling Survey

TING YAN, NORC - University of Chicago, United States
(co-authored with Datta Rupa)

Surveys employing an Address-Based Sampling (ABS) design are gaining popularity among survey organizations and survey researchers in the recent years. For one thing, ABS surveys provide a better coverage of general population than RDD surveys. Secondly, ABS surveys have the flexibility of contacting and recruiting respondents via multiple modes of administration over both RDD surveys and dual-frame surveys with a cell-phone frame. Given that respondents react differently to contacting and recruiting by different modes and different modes of data collection are subject to different types of measurement error, it is time to examine ABS surveys from a total survey error perspective, to evaluate the trade-offs between nonresponse and measurement error by modes of data collection, and to track changes in total survey error by modes.

The 2010 Census Integrated Communications Program Evaluation (CICPE) provides a unique research opportunity to study total survey error in an ABS survey. CICPE is a survey conducted by NORC at the University of Chicago to evaluate the effectiveness of the communications campaign launched by the Census Bureau to promote participation in the 2010 Census. The survey asks about people’s knowledge about and attitudes towards the Census, exposure to various components of the Integrated campaign, and people’s intent to mail back the Census form and whether or they mailed back the Census form.

CICPE was designed as a mixed-mode study using an ABS sample. Selected sample addresses were first matched with telephone numbers. Those that were successfully matched with a telephone number were first contacted in the phone shop by telephones.

Nonparticipants in the phone survey and sample without matched phone numbers were approached to complete an in-person interview. In addition, the Census Bureau provided the actual Decennial census behavior for our sample. In other words, for every sampled household in CICPE, we have their self-report on whether or not they mailed back the Census form and their actual Census form return status (whether or not they returned a Census form and when). Thus, for this variable of interest, we will be able to estimate both measurement bias and nonresponse bias. We plan to study the total bias (the sum of measurement and nonresponse bias) as a function of sample characteristics, mode of data collection, and/or sample progress status.

Using Measurement Models to Locate the Sources of Survey Mode Bias

THOMAS KLAUSCH, Utrecht University and Statistics Netherlands, The Netherlands
(co-authored with Barry Schouten and Joop Hox)

The survey mode acts simultaneously as a relative selection mechanism and a measurement instrument in the creation of mode bias. But how should we usefully describe (or model) mode impact on measurement bias? And how can relative mode-specific selection effects be included in a usefully chosen model, and, subsequently, be adjusted for?

Regarding the first question, we suggest that measurement bias can be studied most usefully after conditioning distributions on true scores, which follows definitions of item bias developed in psychometrical research (Mellenbergh, 1989; Meredith, 1993). By conditioning on true scores it is assured that the answers given by respondents, who are identical with respect to the construct of interest, are compared across modes. While in the majority of cases true scores are unavailable, they can be sometimes conceptualized as 'latent variables' (Borsboom et al., 2003). It is these cases that we focus on in this paper and seek to solve by means of measurement models.

Measurement models, e.g. factor models, describe relationships between postulated 'latent' and 'observed' variables. Mode bias can be defined as the event when a measurement model 'functions' differently between survey modes (Millsap, 2011). Multi-group factor analysis models allow locating mode bias in different elements of the response functions which 'map' latent scores on observed variables. This includes, for example, whether mode differences in observed variable (co-)variances are caused by different relationships of the latent score and the observed score or by simple random error, or the question, if answer categories are used in the same way by respondents in different survey modes.

Our general idea is to estimate the same measurement model in each mode group and consequently test for parameter equivalence in a sequence of steps, which is often referred to as 'measurement invariance testing', where invariance denotes absence of bias. This has only seldom been tried for survey modes (but see de Leeuw et al., 1996; Heerwegh & Loosveldt, 2011). We will apply factor models for ordered categorical data (Muthén, 1984) due to the ordinal scale level of many survey target variables.

Regarding the second question, we argue that it is uncommon to account for relative selection effects in measurement models (beyond allowing for differences in means and variances of latent variables between groups). We show that not accounting for selection can, however, affect the correctness of measurement invariance assessment, i.e. equality tests of parameters like loadings and thresholds, if measurement bias persists on variables that are relevant for selection into survey modes. We made a simulation illustrating this sort of bias in multi-group ordinal factor analysis, showing that inverse propensity weighting (e.g. Rosenbaum, 1987) is the most effective way to adjust for the

relative selection and to yield unbiased measurement invariance tests given the selection variables.
We go on presenting an empirical application, in which a set of items from the European Social Survey, which has been shown in pretests to form a two-dimensional scale, was measured under four different survey modes (CAPI, CATI, Mail, and Web). We adjust for selection – as possible by the available auxiliary variables – and illustrate where differences in measurement emerged by applying the steps of measurement invariance testing, in which the different parameters of the measurement models are constrained equal and changes in model fit are assessed (e.g. Millsap, 2011). For four survey modes such proceeding is complex, because on each step (e.g. equal loadings, or equal thresholds) some modes may be equal while others differ. Our general hypothesis in this respect is that the interviewer modes, CAPI and CATI, will be measurement invariant, as will the self-administered modes, Mail and Web, be (De Leeuw, 1992). We expect to find differences, however, between these interviewer and self-administered mode groups, because of chief differences in measurement processes (such as visual vs. audible information transmission or the presence of an interviewer).
These differences are smaller or absent when comparing CAPI with CATI or Mail with Web (de Leeuw, 2008), which is why we do not expect bias in these comparisons.

We intend to discuss with workshop participants the implications of our findings for the (in)comparability of measurements taken in different survey modes. If our hypothesis is confirmed it appears unrealistic for this set of items to compare measurements between interviewer and noninterviewer modes or to even combine samples in single data sets, because a categorical measurement in CAPI or CATI would have a substantially different meaning compared to Mail or Web. That is, two given respondents, each from a different mode group, but with the same latent trait score would have differing response probabilities across the categories of ordinal indicators.

Finally, we speculate that measurement models could be used to adjust for relative measurement bias. From the observed score of any respondent his/her expected true score under a given mode could be estimated, from which, using the parameter estimates from a second mode group, the expected observed score under the second mode could be derived. This transformation, however, would perhaps only work for a given set of items, a fixed population, under the assumptions of confirmatory factor analysis models and absence of hidden selection bias.

Study of Mode Effects in an Embedded Experiment

PETER LUNDQUIST, Statistics Sweden

A single-mode design is compared with a mixed mode design in an embedded experiment. The single-mode design is a mail survey, while the respondent in the mixed mode design could choose between: mail and web. The ordinary design, the single-mode, is compared with the new mixed-mode design. The ambition is to find a design that produces better quality in terms of accuracy and/or is more cost efficiency.

In the investigation two aspects are considered: Is it possible to use both the single-mode subsample and the new mixed-mode subsample in the statistical production and does the new design give a better quality.

An example on how to incorporate the mixed-mode design in the estimation is given. To evaluate the effect of the new design nonresponse effects and design effects for central target variables are studied. In the evaluation also the selection effect for the mail survey is studied. Nonreponse errors as well as measurement errors are studied.

An Evaluation of Three Surveys Among Non-Western Minorities in the Netherlands with Respect to Nonresponse and Measurement Error

JOOST KAPPELHOF, The Netherlands institute for Social Research, The Netherlands

In recent years the Netherlands institute for Social Research/SCP has conducted three surveys among ethnic minority populations in the Netherlands. These surveys varied in the manner in which they employed general and tailor made response enhancing measures such as the number of contact attempts, translated questionnaires, re-issuing refusals and bi-lingual interviewers. Also other design features differed from survey to survey. For instance one of these surveys was a sequential mixed mode design while the others were single mode face-to-face surveys. The interest is in which set of design features leads to the sample that best reflects the target population. One way to determine the relative success of these different measures and design features is the achieved response rate of a survey, but a higher response rate does not necessarily mean a more balanced sample. The latter can possibly be determined with the use of the R-indicator (representativity indicator) that evaluates the final response composition of the sample with respect to several auxiliary variables (Cobben & Schouten, 2007; Schouten, Cobben & Bethlehem, 2009).

Another relevant question has to do with measurement. In which way do all these different response enhancing measures and design features influence the measurement of survey outcomes? Especially the use of different modes seems to be an important cause of differences in measurement. Recently a method has been developed that makes it possible to disentangle mode and selection effects in a sequential mixed mode survey (Vannieuwenhuyze et al., 2010; Vannieuwenhuyze & Molenberghs, 2010; Vannieuwenhuyze et al., 2012).

This presentation will discuss the representativity of three surveys among non-western minorities (SIM2006, SIM2011 face-to-face and SIM2011 mixed mode) using the R-indicator. Furthermore, it will discuss the measurement differences caused by using different modes by employing the method developed by Vannieuwenhuyze et al.(2010).

Adaptive Survey Designs that Minimize Nonresponse and Measurement Risk

BARRY SCHOUTEN, Statistics Netherlands
(co-authored with Melania Calinescu)

Following cost constraints and technological advances, in recent years, a strong focus on methods for survey data collection monitoring and tailoring has emerged as a new paradigm to efficiently reduce nonresponse error. Paradata, responsive survey designs and adaptive survey designs are key words in these new developments.

In most surveys, all sample units receive the same treatment and the same design features apply to all selected people and households. Adaptive survey designs originate from the idea that different households or persons may respond differently to various design features. Partial R-indicators may be used to identify promising survey designs, and historic survey data is employed to estimate response probabilities given registry data, frame data and paradata.

To date, literature on responsive and adaptive survey designs has concentrated on nonresponse error. In multi-mode survey designs and panel studies, the restriction to nonresponse error is too limited, and one needs to consider measurement as an additional source of error. The extension of adaptive survey designs to measurement error is, however, not straightforward as this type of error is specific to a survey item.

In the presentation, it is discussed how survey designs may be tailored to optimize response rates and representativity using adaptive survey designs. We also sketch a number of approaches how to extend the framework to measurement error. We illustrate the approaches using the Dutch Labour Force Survey.

How Much Do Planned Missing Designs Increase Survey Error In Longitudinal Panel Studies?

DAVID R. JOHNSON, The Pennsylvania State University, United States
(co-authored with Veronica Roth and Rebekah Young)

A Planned missing (PM) design can be a useful tool for reducing both respondent burden and the cost of data collection. By excluding parts of the survey instrument for randomly selected survey participants, and then using modern methods to handle the missing data generated by the design, the statistical parameters estimated will be unbiased but are also likely to be less efficient (Johnson, Roth and Young 2011). Recently, we examined the consequences of a PM design for survey error for scales of health-related behavior in a large national cross-sectional survey which used a PM design to reduce the length of each scale by one-third. Our research found that the scales with randomly deleted set of items performed well in terms of scale means, distributions and covariances with predictive characteristics when the dropped items were imputed using multiple imputation. How well these measures perform in a longitudinal panel study, and how much survey error is introduced by a PM design, remains unclear. There is little systematic research evaluating the consequences for survey error of PM designs incorporated in longitudinal panel survey studies. There has been some work on PM designs in multiple wave longitudinal studies using simulations to assess the specific planned missing design that can yield the most efficient estimators for growth curve models (e.g., Graham 2001). To our knowledge, there is no research on the consequences of PM designs that empirically assesses the reliability, stability, and efficiency of this strategy in a panel study. The purpose of our study is to contribute to this literature by exploring the performance of health related scales in a nationally representative two wave panel survey.

Our study analyzes three health-related scales included in the National Study of Fertility Barriers (NSFB) (Johnson et al. 2009). These are the CES-D (depression) scale (Andresen et al., 1994), a Medical Locus of Control Scale (Wallston et al., 1978), and an 8-item scale constructed for this study that assessed respondents' attitudes about the ethics of infertility treatments (Ethics of ART). The NSFB is a nationally representative telephone survey of 4,700 women age 25 to 45 (and their available partners) with the baseline survey conducted in 2005-2008 and a three-year follow-up survey completed in 2009-2010 with 60% of the wave 1 respondents.

A PM design was included for 20 scales assessed in the surveys. In this design, each scale was divided into three sets of items. A random number was drawn in Wave 1 for each respondent for each scale to determine which of the three sets would be dropped. There was also a small fraction of the respondents who were randomly selected to receive all the items in the scale. Two types of scale scores were created for each respondent; the first was based on the mean of all available items and the second imputed the items dropped in the PM design which were summed along with the observed items. In the second wave most respondents were administered the same set of items they received in wave 1.

Using both waves of data, our analysis compares the reliability, stability, and standard errors of the scales among those who did and did not receive the PM design. We use these findings to assess the PM design's impact on bias and random error. Based on these findings we draw some recommendations about the advantages and tradeoffs involved in using PM designs in longitudinal panel studies.

Planned Missing-Data Designs and Statistical Matching: A Smart Response to Minimising Total Survey Error?

FEMKE DE KEULENAER, Gallup Europe, Belgium
(co-authored with Robert Manchin)

In survey research, reaching a sufficient level of precision requires a large enough sample size and a detailed measurement instrument. Large samples, however, are expensive to obtain, and lengthy questionnaires can result in increased non-response and respondent burden, modifying answering behaviour and, eventually, increasing measurement error. In the past few years, at times of declining response rates and rising survey costs, planned missing-data survey designs, combined with appropriate statistical matching and imputation techniques, have received increasing attention as a tool to improve surveys by reducing total survey error.

Increasing response rates and controlling respondent burden are important challenges in the Gallup World Poll, a large-scale cross-national survey that has been conducted annually since 2006. In this study, a planned missing-data survey design could be implemented, for example, to control the length of the interview – and, as such, could lower response burden, increase respondent engagement and potentially also reduce total survey error. In a missing-data design, all respondents are asked the same set of core questions, but different respondents will be presented with different additional subsets of questions (e.g. a detailed employment module vs. an instrument to measure health conditions).

Using a missing-data design implies that, at the end of the data collection period, researchers have two or more separate data sets/surveys to analyse. Statistical matching, however, can be used to generate a new synthetic data set that allows more flexible analysis than would be possible with separate data sets. In statistical matching, variables common to each data set are used to identify similar respondents that can be matched to create the synthetic data set. In this paper, constrained statistical matching based on estimated propensity scores will be used in an attempt to match different subsets of the Gallup World Poll.

Although planned missing-data designs appear to be a smart response to classic designs, using missing-data designs without appropriate statistical tools may do more harm than good. Statistical matching procedures are inferential procedures based on suitable statistical models (e.g. regarding the relationship among the variables to be matched); incorrect specification of such models undermines the overall quality of the final result. In this paper, we will not only look at the potential advantages of planned missing-data designs and statistical matching to reduce total survey error in the Gallup World Poll, but we will also study potential drawbacks of this approach in terms of a loss in precision and power of the analysis.