Final Abstracts in Order of Presentation

Sunday, September 20, 2015
9:30-11:30 a.m.
Paper Session I

Interactions of Survey Error and Ethnicity I
Session Chair: Sunghee Lee

Invited Presentation: **Ethnic Minorities in Surveys: Applying the TSE Paradigm to Surveys Among Ethnic Minority Groups to Assess the Relationship Between Survey Design, Sample Frame and Survey Data Quality**

Joost Kappelhof[1]
Institute for Social Research/SCP[1]

Minority ethnic groups are difficult to survey mainly because of cultural differences, language barriers, socio-demographic characteristics and a high mobility (Feskens, 2009). As a result, ethnic minorities are often underrepresented in surveys (Groves & Couper, 1998; Stoop, 2005). At the same time, national and international policy makers need specific information about these groups, especially on issues such as socio-economic and cultural integration. Using the TSE framework, we will integrate existing international empirical literature on survey research among ethnic minorities. In particular, this paper will discuss four key topics in designing and evaluating survey research among ethnic minorities for policy makers. First of all, it discusses the reasons why ethnic minorities are underrepresented in survey. In this part an overview of the international empirical literature on reasons why it is difficult to conduct survey research among ethnic minorities will be placed in the TSE framework. Secondly, it reviews measures that can be undertaken to increase the representation of minorities in surveys and it discusses the consequences of these measures. In particular the relationship with survey design, sample frame and trade-off decisions in the TSE paradigm is discussed in combination with budget and time considerations. For instance, the advantages and disadvantages of different data collection modes (face-to-face, telephone, web, postal or sequential mixed mode) and response enhancing measures such as the use of translated questionnaires (including best practices), bilingual interviewers and interviewers with a shared ethnic background, and how this, for example affects the trade off between measurement error and nonresponse error. This will be illustrated by empirical research based on a large scale project on surveying among ethnic minorities in the Netherlands. The third part discusses the empirical literature on different methods that can be applied to assess the data quality of surveys among ethnic minority groups and how this can be utilized to assess the representation and measurement of ethnic minorities in national surveys. The fourth part deals with potential sources of method bias that can arise as a result of survey design choices in surveys among ethnic minority groups. In particular how this can affect the cross cultural comparison of survey results when surveying different ethnic minority groups. This part will draw on existing international literature about cross-cultural survey research and best practices, such as the CSDI initiative (Survey Research Center, 2010) as well as empirical research based on a Dutch large scale project on surveying among ethnic minorities.. This chapter will conclude with lessons learned on surveying ethnic minorities in the Netherlands and discusses how these lessons are relevant for an international audience.

How Does Language Shape Public Opinion?

Efren Perez[1]
Vanderbilt University[1]

Growing evidence suggests that public opinion varies by language of interview, yet modest theory exists to explain this pattern. I propose a theoretical framework where language affects people's opinions by conditioning the mental

accessibility of relevant political concepts. I claim that political concepts vary by how associated they are with certain languages, which means that people stand a higher chance of acquiring a construct when it is tied more to the tongue one speaks. Moreover, I argue that recalling political concepts from memory is easier when the language a construct is linked to matches the tongue one interviews in. Combined, these elements should accentuate the opinions people report by increasing the accessibility of some concepts. I test my theory by manipulating interview language in a U.S. survey of English/Spanish bilingual Latinos. I find that English interviewees report higher levels of opinions based on political concepts that are more associated with English (e.g., political knowledge), but lower levels of opinions based on concepts that are more connected to Spanish (e.g., Mexican identity). I then rule out that these effects arise because a) my survey items are incomparable across languages; b) respondents become exceedingly emotional (e.g., anxious) when interviewing in a minority tongue (i.e., Spanish); and c) interviewees feel more politically efficacious when interviewing in a dominant language (i.e., English). I conclude by discussing the import of my results for research in multilingual settings.

Survey Error Associated with the Translation of U.S. Education Level Categories into Spanish

Patricia Goerman[1], Leticia Fernandez[1], Kathleen Kephart[1]
U.S. Census Bureau[1]

Questions about educational attainment are difficult to translate for use with respondents from different countries. This is particularly the case for Spanish-speakers in the United States, who come from a variety of countries where educational systems are different not only from the U.S. system but from each other as well. A cognitive testing study of the 2009 Spanish-language version of the American Community Survey (ACS) found that Latin American immigrants with limited English proficiency residing in the United States may have been misinterpreting educational attainment question categories on Census surveys, potentially resulting in higher response error and a reduction in data quality. For example, Mexican-origin respondents interpreted "Diploma de escuela secundaria," the original translation for "Regular high school diploma," to correspond to nine years of schooling. Similarly, the translation for "Bachelor's degree" ("Titulo de bachiller universitario") was interpreted appropriately by Puerto Rican Spanish speakers, but not by respondents from Argentina, Mexico, Colombia and Nicaragua. In these countries, the term "bachillerato,"which sounds very similar, is used to describe either junior high or high school. Both of these translations could have biased the measurement of immigrants' educational attainment since they led respondents to report lower levels of education as higher ones. As a result of this cognitive testing study, in 2011 several minor changes were made to the high school and bachelor's degree translations in the Spanish ACS education question. The translation for Regular high school diploma was changed from "Diploma de escuela secundaria" to "Diploma de escuela secundaria o preparatoria"; the translation for Bachelor's degree was changed from "Titulo de bachiller universitario" to "Titulo de licenciatura universitaria"; and the translation for Professional degree beyond a bachelor's degree was changed from "Titulo profesional mas allá de un titulo de bachiller" to "Titulo profesional mas allá de un titulo de licenciatura universitaria." While these changes should improve data quality, we suspect that the issues will not be completely resolved, since terms that can have multiple meanings depending on country of origin are still included. In this paper, we compare the responses of Latin American immigrants with limited or no English proficiency to the educational attainment question in ACS before and after implementation of the translation changes. To assess the impact of these changes, we use two data sets: the ACS 3-year files from 2008-2010 (before the changes were implemented) and from 2011-2013 (after the changes took place). If changes in wording resolved the interpretation issues, we should see an increase in the percentage of respondents placing themselves in the grade 9 category rather than at the high school level, and a shift towards reporting a bachelor's degree rather than a professional degree. We examine these distributions by length of residency in the United States and reported English proficiency. We hypothesize that more years in the U.S. and higher English-speaking proficiency will result in greater familiarity with the U.S. education system. Preliminary results support these hypotheses. We end with a discussion of possible next steps to evaluate and improve education level translations.

The Role of Time Perspectives on Subjective Probability Response Patterns Among Hispanics

Sunghee Lee[1]
University of Michigan[1]

Questions using subjective probability have gained popularity increasingly in recent years. These questions ask about respondents' perceived chances of various future events. Popularity of subjective probability questions is due to empirical evidence□answers to these questions do predict actual future outcomes and behaviors. An implicit yet important premise of subjective probability questions is that respondents have stored cognitively organize their personal experiences to be relevant for the future, theorized as future-oriented time perspective. Time perspective is an unconscious yet fundamental cognitive process that provides a framework for organizing personal experiences in temporal categories of past, present and future. With Hispanics described to be present oriented, this premise of subjective probability questions does not align. This paper hypothesizes that the difficulty of these questions is higher for Hispanics as a group than Whites, consequently leading to higher item nonresponse rates and that independent from the race/ethnicity, past- or present-oriented individuals are associated higher item nonresponse. Using the data from the Health and Retirement Study, we

examine this measurement issue that affects nonresponse error.

Sexual Orientation and Behavior Item Nonresponse and Measure Concordance Across Race/Ethnicity, Sex, Interview Language, and English Proficiency: Evidence from Five Cycles of the California Health Interview Survey

Matt Jans[1], David Grant[2], Joseph Viana[2]
Center for Health Policy Research[1], UCLA Center for Health Policy Research[2]

Research on the lesbian, gay, and bisexual (LGB) population uses three measures of "sexual orientation." Sexual identity questions ask respondents which orientation they identify with most. Response options usually include "gay, lesbian, bisexual, and something else." Sexual behavior questions ask respondents to report the sex of their sexual partners over some timeframe. Response options commonly include "only men, only women, and both men and women." Sexual attraction questions ask respondents about which sex(es) they are attracted to Response options commonly include "mostly men, mostly women, or both men and women." These questions address different components of sexuality, but sometimes they are conflated in sexual health research constituting a validity error under the TSE framework (Groves et al., 2009). Validity errors happen when the measure used does not reflect the construct the researcher intended to measure. They may also create coverage error if the question is used for screening during recruitment or sub-setting in analysis. For example, inferring sexual identity from the sex of a respondent's sexual partners in the past 12 months would mis-represent the identity of respondents who have been celibate over that timeframe. Lifetime sexual partners misclassifies respondents who have not yet had a same-sex sexual experience, even if they experience same-sex attraction or identify as LGB. Sexual orientation is a complex and fluid concept, so respondents with same-sex sexual experiences may not answer sexual identity and sexual attraction questions in obviously-consistent ways. This is reflected in CDC's use of "men who have sex with men" instead of "gay" or "homosexual" in HIV/AIDS prevention material and research. This study tackles this issue from a TSE perspective by focusing on item nonresponse rates to sexual identity and past-year same-sex sexual activity as asked in the California Health Interview Survey (CHIS), and calculates differences in reported rates of LGB classification from each of those two questions. The sexual identity question reads, "Do you think of yourself as straight or heterosexual, as gay or homosexual, or bisexual?" The sexual behavior question reads, "In the past 12 months, have your partners been male, female, or both male and female?" Respondents age 18-70 years from five cycles of CHIS will be used (n = 182,812) to evaluate two outcomes: a) the item nonresponse rate to each question (identifying "don't know" and "refusal" responses separately), and b) the concordance between LGB classification based on each question. Hispanics are more likely than Whites to have missing sexual orientation when ethnicity alone is considered, but the missingness may be due more to interview language. Based on 2009 data, we find that about 73,000 population members (i.e., weighted respondents) have "incongruous" responses (e.g., straight men who report sex with men). Most of this "error" comes from straight-identified respondents reporting sex with same-sex partners, not gay or lesbian respondents reporting sex with opposite-sex partners. We will investigate whether Hispanics or non-Hispanics are more prone to errors of validity when researchers use sexual behavior as a naïve measure of sexual orientation.

Interviewer Effects Across Error Sources

Session Chair: Stephanie Eckman

Interviewers Attitudes Towards Surveys, Interviewer Experience, Response Locus of Control, Personality and the Impact on Cooperation with CATI-Surveys

Volker Hufken[1]
Heinrich-Heine-Universität Düsseldorf[1]

We examine the effects of interviewers' experience, interviewers co-operation related locus of control, attitudes towards surveys, and personality traits on interviewer performance in terms of the likelihood of refusal in a cross-sectional telephone surveys. Previous studies of the association between co-operation and interviewer skills and interviewer attitudes have not directly addressed the role on CATI-surveys. We use 250 interviewers and analyse co-operation outcomes for over 15,000 cases. We find evidence of effects of experience, attitudes and personality traits (e.g. extraversion) on co-operation. The implication for survey practice for example, it might be useful to administer a personality test and obtain the applicant's score on the extraversion dimension.

Mode Effect in Voting Behavior and Social Trust: A Comparison Between CAPI and CATI

Pei-shan Liao[1]
Academia Sinica[1]

Response distribution and data quality are influenced by different mode of data collection, which is seen as mode effect. One significant influence of mode effect on data quality is social desirability bias, which is related to whether an interviewer is involved, pace of cognition process during interviews, sensitivity of survey questions, etc. For example, social desirability bias is more likely to occur in face-to-face interview when compared to telephone interview. Slower pace and the development of rapport in the former encourage respondents to think thoroughly and provide a socially desirable answer. However, previous studies on social desirability of mode effect did not obtain consistent findings. The problem of social desirability bias is found to be more severe in telephone interview than in face-to-face interview when dealing with voting turnout. The issue of such responding bias with the consideration of mode effect is worth further examination. This study aims to examine mode effect on social desirability bias by comparing telephone and face-to-face interviews, mainly on the voting turnout of the 2012 president election in Taiwan and social trust. Data are drawn from two national representative sample conducted in the summer of 2014 to eliminate recall error. Data of telephone survey has been collected using CATI resulting in a total of 3,379 complete cases. The face-to-face survey data, which are available in the end of 2014, are drawn from the Taiwan Social Change Survey, module of citizenship collected using CAPI. Both samples will be weighted by population characteristics. Socio-demographic variables will be compared first to examine the dis/similarity of the two samples. Voting behavior, social trust and socio-demographic variables will be included in the multivariate analysis. Conclusion and discussion will be provided.

Influence of Prior Respondent-Interviewer Interaction on Disclosure in Audio Computer-assisted Self-interviewing (ACASI)

Hanyu Sun[1]
Westat[1]

Audio computer-assisted self-interviewing (ACASI) is one of the best methods for collecting information about sensitive topics such as illicit drug use or sexual behavior. In an ACASI interview, respondents read questions on a computer screen and simultaneously hear the text of the questions read to them through headphones. Many studies have found that ACASI increases sensitive disclosures relative to other methods, such as computer-assisted personal interviewing (CAPI) and paper-and-pencil self-administered questionnaires (e.g., Tourangeau & Smith, 1996). According to conventional thinking, ACASI is taken as an independent mode of data collection, i.e., the CAPI interaction that almost always precedes it is rarely considered when assessing its impact on disclosure. However, none of the existing research has investigated the possibility that the interviewer-respondent interaction in the prior CAPI module may affect disclosure in ACASI. The prior interviewer-respondent interaction may create a sufficient amount of social presence to reduce sensitive disclosures in ACASI. The respondent may have built a positive relationship or rapport with the interviewer during their prior interaction. Additionally, if the voice used in the ACASI audio-file sounds similar to the CAPI interviewer, it may work as a reminder of the presence of the interviewer. It is plausible that more social presence, created in the preceding module (CAPI or video-mediated interviews), may lead to fewer sensitive disclosures in the ACASI module. We test this carryover effect with a laboratory experiment to see whether the interaction between the interviewer and the respondent in the preceding module (CAPI or video-mediated interviews) has an effect on the reporting of sensitive information in a subsequent ACASI module. Eight professional interviewers and 128 respondents participated. We found no significant difference on disclosure between the same voice and the different voice condition. However, there were marginally significant carryover effects of rapport in the preceding module on disclosure in the subsequent ACASI module. Respondents who experienced high rapport in the preceding module gave more disclosure in the subsequent ACASI module. Furthermore, compared with ACASI, the percentage of reported sensitive behaviors was higher for video-mediated interviews for some of the highly sensitive questions.

Nonresponse

Session Chair: Lin Wang

Identification and Reduction of Nonresponse Bias in Address-Based Sample Surveys

Burton Levine[1]
RTI[1]

Dual-frame random digit dialing (RDD) telephone data collection and address-based sampling (ABS) with a mail contact are two commonly used probability survey designs. Data from these different study designs may produce different

population estimates for the same outcome due to different sources of survey error. In 2014, the New York Adult Tobacco Survey (NY-ATS) was fielded both as a dual-frame RDD telephone survey and as an ABS survey with mail contact. The two studies produced substantively meaningful different smoking estimates as well as statistically significant differences in other study outcomes. I estimated smoking prevalence for New York block groups by fitting a model using the previous 4 years of NY-ATS data. I found a significant correlation between estimated smoking propensity and response propensity in the ABS sample, but not in the listed landline sample, thereby, demonstrating nonresponse bias in the ABS study. I modified the weighting procedure for the ABS data to use the block group estimated smoking propensity in the nonresponse model. The adjustments based on this model reduced the nonresponse bias in the ABS sample and reduced the difference in the smoking estimates between the ABS and RDD studies. Similar methodology can be applied to studies with outcomes other than smoking.

How to Estimate the Non-response Error in Probability Terms

Daniel Thorburn[1]
Department of Statistics, Stockholm University[1]

One problem with quality declarations is that the quality of the different error sources is measured on different scales and that these measures cannot be added together into one measure. Bayesian methods often make this easier. In this paper we show how the non-response error can be formulated so that it can be merged with the sampling error into one measure, e.g. 95 % prediction intervals. To describe the idea this abstract starts with a simple example. Suppose that we have conducted a one-variable sample survey with a large non-response and also that we in the frame (or register) have access to 19 other background variables. Calculate the loss of error on these 19 variables for the sample compared to the frame values. Assuming that these 19 variables and the survey variable can be considered similar we may assume that the 20 non-response errors are exchangeable, i.e. we have no reason a prior to believe that one is larger than the other. Thus we can conclude that with probability 95% the survey error does not exceed the largest of the other 19 errors. This idea will be worked out in more detail using statistical distributions like the chi -2- distribution in order to be able to handle any (positive) number of auxiliary variables and coverage probabilities. Since this will give a standard type of probability distribution it can easily be combined with ordinary random sampling errors to a total. In practice, when one has auxiliary variables in the frame these are often used in the estimation stage, e.g. by some type of imputation or calibration/propensity score weighted to reduce failure error. The above procedure will not work then, since we do not know how much of the random error that has been removed and we have no new auxiliaries to compare with. But then we instead can proceed like this. First we predict each of the 19 variables with the other 18 as background variables using the same technique as well as possible. Then we have 19 non-response adjusted estimates with non-response errors and one non-response adjusted estimate from the survey. We may thus draw the same conclusion as we did above, i.e., with 95% probability the error of the study variable is not the largest. Also this will be put into a Bayesian statistical framework in order to handle any number of auxiliaries. We also discuss the possibilities to include other prior knowledge into the calculations. The paper will include a practical example.

An Imputation Approach to Handling Nonignorable Nonresponse Using Paradata

Seho Park[1]*, Jae Kwang Kim*[1]*, Kimin Kim*[2]
Iowa State University[1]*, Korea Labor Institute*[2]

Paradata is often collected using the survey process to monitor the quality of the survey response. One such paradata is the respondent behavior, which can be used to model the response propensity. Instead of the usual nonresponse adjustment method (Kott, 2006) based on this paradata, we propose an imputation approach, which can significantly reduce the nonresponse bias and provide more efficient survey estimates. The proposed method is particularly useful when the response mechanism is nonignorable without this paradata, but becomes ignorable with this paradata. The proposed method is applied to Workplace Panel Survey in Korea.

A Forest Full of Respondents: Comparing Logistic Regression and Random Forest Models for Response Propensity Weighting Adjustments

Trent Buskirk[1]*, Stanislav Kolenikov*[2]
Marketing Systems Group[1]*, Abt SRBI*[2]

Response rates for modern surveys are trending downward leaving the potential for nonresponse bias in the resulting estimates. Nonresponse may be a complex function of known auxiliary variables or latent variables not typically measured or not available on the sampling frame. The magnitude of this nonresponse bias can be reduced by using propensity weighting adjustments obtained from models that estimate survey response such as logistic regression. However, for

smaller samples with many survey response predictors or when survey response is a complex function of many variables that interact, logistic regression may fail to converge because of empty cells or perfect separation. In this paper we compare logistic regression to a newer machine learning technique – random forests – for estimating response propensity. Random forests is an ensemble method that assigns final estimates of survey response by aggregating estimates across a collection of classification trees and is well suited across a range of sample sizes and number of predictors. In this research we explore the utility of nonresponse adjustments that are based on survey response propensities estimated using both logistic regression and random forest methods. The sample was selected from a realistic finite population comprised of a subset of data from the U.S. National Health Interview Survey and contains information on a set of 13 demographic, household and health related variables serve as predictors in the propensity models. Survey response outcomes were simulated for each member of the sample according to both a simple and complex response mechanism. Final sampling weights were computed using both direct response propensity as well as propensity stratification adjustments using the estimated response propensities. The total error of weighted survey estimates for five key outcomes is evaluated by incorporating estimates of both bias and variance. Design effects and overall model fit statistics are also presented for both methods. Compared to logistic regression, estimated response propensities from random forests are less correlated with actual propensities generated under the simple response mechanism. On the other hand, when the survey response mechanism is complex, random forests appear to offer marginal improvements in survey estimates over logistic regression using direct propensity adjustment. Propensity stratification gave surprising results across both response mechanisms.

The Intersection of Sampling and Nonresponse: Does Repeated Sampling of Some Individuals Affect Nonresponse Bias?

Jennifer Sinibaldi[1], Anton Örn Karlsson[2]
JPSM, University of Maryland[1], Statistics Iceland[2]

In the interest of better understanding total survey error, survey researchers are increasingly analyzing multiple stages of the survey lifecycle to understand how early stages might impact later ones. To contribute to this effort, this analysis examines how the repeated sampling of individuals affects nonresponse bias.
Although government surveys of the general population in populous countries are not likely to sample the same individual repeatedly, this is a possibility when the population is small. Iceland has 325,000 people but four major household surveys, including the Labor Force Survey. Combining all of the sampled cases from 2002 to 2013 for these four surveys plus another household survey finds that almost 11,000 individuals have been selected for more than one household survey during this period. To determine if the repeated survey invitation affects response, we estimate the likelihood of the twice-sampled case to respond to the second survey invitation compared to the response likelihood of the cases only in one sample. Upon identifying a different response likelihood for the twice-sampled cases, we use rich auxiliary data from the national registry to identify characteristics that may suffer from nonresponse bias. We also isolate these cases to study the effect of time between survey invitations.
While Iceland is a somewhat unique case, the lessons from this analysis are broadly applicable. When all types of agencies that conduct surveys (e.g. academic, polling, market research) are considered, some individuals will be sampled twice, even within large populations. The results of this analysis will allow for conservative estimates of the effect on response rates and nonresponse bias for individuals who are sampled twice across different agencies.

Comparison Between Substitution and Strata Collapsing for Sampling Variance Estimation Under the Presence of Nonresponding PSUs

Raphael Nishimura[1], James Lepkoswki[1]
University of Michigan[1]

In stratified cluster sample designs with few Primary Sampling Units (PSUs) per stratum there is a high risk that some of the strata will end up with only one or none responding PSU after nonresponse, which poses a problem for sampling variance estimation. When this occurs, a common strategy to estimate sampling variability measures is to form pseudo strata with at least two PSUs by collapsing strata with one or none PSU. Such procedure, however, tends to overestimate sampling variability. An alternative approach is to substitute the nonresponding PSUs by units that were not originally selected in the sample. Although such procedure has been extensively criticized in survey sampling literature, it is largely used in practice, especially in school-based surveys. Moreover, one of the possible advantages of substitution, according Vehovar (1999), is that it maintains the sample design structure, so that, if substitution is fully successful, it is possible to perform sampling variance estimation using standard techniques. However, as Vehovar (1999) also points out, it is still necessary to compare these two approaches in term of the mean square error of their sampling variance estimates. In this paper we conduct such comparison using a large-scale simulation study.

Data Falsification: Who, Why and How to Detect It

Session Chair: Nancy Bates

Taking Fabrication Detection and Prevention Beyond the Interviewer Level

Steve Koczela[1]
The MassINC Polling Group[1]

Most of the literature on survey data fabrication focuses on preventing and detecting "curb-stoning", or data fabrication at the interviewer level. While curbstoning has been shown repeatedly to be a problem in need of sustained attention, new evidence suggests that fabrication often happens during other phases of data collection and processing. Fabrication prevention and detection methods need to evolve to account for the possibility of fabrication beyond the individual interviewer. Survey data often passes through multiple hands on the way to publication, including interviewers, supervisors, keypunchers, and managers, among others. Each of these levels presents its own problems in preventing and detecting data fabrication. Methods of analysis that reveal a dishonest interviewer may not detect fabrication from other staff. Without methods to prevent and detect fabrication from employees other than interviewers, the risk posed by fabricated data increase significantly. Fabrication higher up the chain can affect a far larger the number of cases than can be tainted by a dishonest interviewer. One example of this larger scale fabrication can be found with analysis of duplicate cases, which this paper explores in detail. Data files from high profile, published surveys have been found to include large numbers of duplicate cases across dozens of variables. They also include "near duplicates", where nearly all variables have been duplicated, but a few numbers have been changed. There are also instances of lengthy strings of duplicated variables. Finally, there are blocks of consecutive cases which are identical to other blocks of consecutive cases, where sets of interviews appear to have been duplicated in different parts of the data file. Each of these instances is extremely unlikely to be attributable to curbstoning by individual interviewers filling out single questionnaires. Someone engaging in classic curbstoning behavior could likely not produce data that fit these patterns, especially on such a massive scale. Instead, some of these patterns suggest fabrication by an individual or group with access to the data file, where the creation of hundreds of perfect or near-perfect duplicates would be as simple as cutting and pasting. Further, these patterns are often concentrated in the data from specific supervisors or specific parts of a survey data file, suggesting the issues extend beyond simple curbstoning. Apparent fabrication beyond the interviewer level means the field of fabrication prevention and detection needs to expand. Statistical detection methods and administrative prevention methods must be developed to counter this threat to data integrity. This paper proposes ideas for improving methods for countering this new threat to data quality.

Curb Stoning and Culture

Arthur Kennickell1
Federal Reserve Board1

Much of the most critical work interviewers perform takes place in a way that is generally only partially or indirectly observable, at best. That weakness makes surveys vulnerable to abuse through data fabrication. This paper considers the nature of data fabrication, its possible detection, and the motives for such behavior. It argues that when observability is so limited and our ability to target incentives toward desirable behavior are also weak, the best hope we have is to foster a culture among field staff that supports the desired behavior.

Data Falsification: Who, Why and How to Detect It

Cathy Furlong[1]
Integrity Management Services, LLC[1]

Cathy Furlong will discuss some of the connections between fraud investigation and curb stoning. The main focus will be: 1. Statistical methods which can be used in both fields. 2. The levels of "organizational structure" in Medicaid/Medicare fraud detection which may also be beneficial to deter curb stoning.

Understanding Core Reasons Why Falsification Occurs During Field Data Collection

Timothy Olson[1]
U.S. Census Bureau[1]

My experience managing field data collection for more than 20 years shows that very little data falsification actually occurs by interviewers. But when it does, why? What are the environmental factors that contribute to data falsification?

Are interviewers evil to the core and simply intent on falsifying data? In a few limited situations, possibly. But most data falsification occurs through a variety of other environmental factors described in this paper. The primary environmental factor that creates data falsification is a misunderstanding of survey procedures or key concepts embedded in survey questions. Interviewers receive training on their survey. In some instances, falsification occurs because they did not fully understand the training. Or the training was substandard. In some cases, the interviewer simply misunderstood what other interviewers comprehended through the training. In these situations, having live-time metadata analytic tools to flag potential misunderstandings of key concepts by the interviewer are critical to avoiding future errors. Another factor that contributes to data falsification is constant pressure by survey management to increase response rates, decrease item nonresponse, and reduce data collection costs. In this environment, interviewers may succumb to the pressure and take short cuts to get a complete interview. Interviewers can streamline how they ask questions, listen to and observe a respondent and then "fill in the blanks" after the interviewer has departed the interview location. In some instances where a personal visit is required, an interviewer might use a phone interview to avoid time and mileage associated with a personal visit. In other situations, especially where there is reluctance to participate, the interviewer will turn to proxies for the required data, often through an adjacent household. Establishing realistic response rate expectations, legitimate cost models for data collection, and training interviewers how to optimize their work are all critical solutions to avoid pressure-driven data falsification. Intensional data falsification by field interviewers does occur, albeit rarely. Simply making up what appears to be a "good interview" while sitting in one's living room certainly does happen. When identified, action must be swift and clear.

Detecting Data Falsification in Survey Research

Noble Kuriakose[1], Michael Robbins[2]
SurveyMonkey[1], Princeton University[2]

Survey research relies on interviewers to accurately and faithfully record the responses that are used for analysis. Unscrupulous in-country firms and field workers have long employed duplicate observations to boost the number of cases delivered to a researcher. Also known as curb-stoning, duplication involves replicating legitimate responses to survey questions and falsely representing the number of interviews conducted. We report that there are also many surveys with near duplicates-- cases where falsified observations have had a small number of variables altered to conceal exact duplicates.Using the treasure trove of publicly available survey data, we establish the degree to which duplicates and near duplicates are present in survey research. We find that international survey work, which is largely done via face-to-face interviews and in the interest of capacity-building, is especially susceptible to this type of fraud. Of the hundreds of surveys we analyzed, about one-third had some presence of near duplicates. About 100 surveys had near-duplicates or duplicates in 50 or more cases. Indeed, duplication affects nearly all of the widely used, major international survey efforts.As part of an effort to establish data quality standards, we build on the long tradition of efforts to ensure quality in survey work and propose a set of interrelated tests that can be employed to help ensure high levels of data quality. First, demographic variables should match the sampling frame. Second, expected correlations should hold between variables, including demographic variables (for example, in Arab countries gender should correlate with attitudinal items on women's role in society). Third, and our main contribution, we argue that the distribution of the duplication (percent of shared responses) between a variable and it's closest kin across the dataset ought to fit a Gumbel distribution with a mean of 0.7 or less and a maximum below 0.9. If all three conditions hold, we argue that data are unlikely to be deliberately falsified.Our paper demonstrates that this problem often goes undetected. In large part, we believe, due to the challenge of detecting these cases. As part of our contribution, we detail a STATA package we wrote to serve as a diagnostic tool and report the incidence of near-duplicates.

Methods for Assessing and Integrating Solutions for Total Survey Error in a Large National Survey: The National Crime Victimization Survey

Session Chair: Marcus Berzofsky

The Impact of Adaptive Design on Nonresponse Bias and Precision

Michael Planty[1], Lynn Langton[2]
U.S. Bureau of Justice Statistics[1], BJS[2]

In survey sample research the balance between achieving high response rates and controlling non-response bias is critically important. Recently the use of adaptive design techniques for nonresponse follow-up has focused heavily on limiting bias and ensuring representativeness. However, the impact on precision is also of concern. Sample surveys are designed to sample an appropriate number of cases to produce a desired level of statistical power. The resulting lack of precision due to nonresponse--nonresponse precision error--is examined here through simulation techniques that assess point estimates and coefficients of variation as potential trade-offs given various strategic nonresponse follow-up

techniques.

Methods for Reducing TSE in Subnational Estimates: An Assessment of Coverage Error Through the Comparison of Nationally Calibrated Weights to Subnational Area-specific Calibrated Weights

Bonnie Shook-Sa[1], Marcus Berzofsky[1], G. Lance Couzens[1], Andrew Moore[1], Philip Lee[1], Lynn Langton[2], Michael Planty[2]
RTI International[1], U.S. Bureau of Justice Statistics[2]

The National Crime Victimization Survey (NCVS), sponsored by the Bureau of Justice Statistics (BJS) and conducted by the U.S. Census Bureau, is a multi-mode, rotating panel survey of households that produces nationally-representative criminal victimization estimates for all major types of crime in the United States. The NCVS has always been a rich source of information about criminal victimization at the national level, but subnational estimates would be useful in better understanding local crime patterns and trends. Of specific interest are estimates in heavily populated states and metropolitan statistical areas (MSAs), as well as "generic area" profiles that provide estimates for geographic areas with similar characteristics in aggregate (e.g. all rural areas in the South with fewer than 100,000 persons). The sample sizes in some large states, MSAs, and generic areas support the calculation of crime estimates directly and with reasonable levels of precision when multiple years of NCVS data are aggregated. However, the NCVS was designed to exclusively produce national estimates, which means analysis weights were created to produce representative victimization counts, rates, and proportions only at the national level, without regard to smaller geographic areas. Thus, subnational estimates could exhibit systematic bias from undercoverage due to variation in the primary sampling units (PSUs) selected within these areas and how the sample was weighted. The national stratification and allocation of the sample does not ensure that selected PSUs within a given subnational area are representative of that area, only that PSUs in aggregate are nationally representative. Therefore, the analysis weights for respondents in these areas represent not only persons in the subnational area but persons in other areas with similar demographic characteristics, as NCVS non-response and post-stratification adjustments do not control weights at the subnational level. We evaluate the NCVS sample in subnational areas and assess the representativeness of key populations and demographic groups. Furthermore, we propose and evaluate weight adjustments to ameliorate coverage concerns within subnational areas. For key estimates, we compare the original and re-calibrated subnational weights to assess the accuracy of the uncalibrated estimates and to evaluate the impact of re-calibration on estimate precision. Our analysis found that, within both states and MSAs in self-representing PSUs, the re-calibration significantly changed the estimates with a relatively minimal impact on precision.

Discussion on Reducing Total Survey Error in the National Crime Victimization Survey

Lynn Langton[1]
Bureau of Justice Statistics[1]

The discussion will focus on tying together the many BJS projects aimed at reducing TSE in the National Crime Victimization Survey (NCVS). It will touch on ways in which findings from these projects could ultimately be integrated into the NCVS design.

Comparing Error Structures for Estimates of Rape and Sexual Assault: The Design of the National Survey on Health and Safety

David Cantor[1], Shannan Catalano[2], Allen Beck[2]
Westat[1], U.S. Bureau of Justice Statistics[2]

Estimates of rape and sexual assault vary widely across different surveys. These differences have resulted in heated debate over the ideal method for collecting self-report data on rape and sexual assault. These disparities have also resulted in confusion as to which estimates are more accurate. This paper describes a study sponsored by the Bureau of Justice Statistics (BJS) that seeks to shed light on this debate and to inform efforts to redesign methods currently employed on the National Crime Victimization Survey (NCVS) to collect data on rape and sexual assault. A recent report by the National Academy of Sciences (2013), funded by BJS, concluded that 'best practice' to collect data on rape and sexual assault should use behavior specific questions (BHQ). BHQ use explicit language to describe the behavior in question (e.g., penetration and sexual touching) and the tactic (e.g., physical force and incapacitation). This approach avoids loaded terms, such as 'sexual assault' or 'rape', which require respondents to impose their own definitions on what happened. However, it is unclear how other design features, such as survey design, coverage/non-response bias, and mode of interview interact with BHQ and affect error. To investigate these issues, Westat, under a cooperative agreement with BJS, is implementing the National Study on Health and Safety (NSHS). The study consists of two phases. The first is to develop and test a methodology that combines BHQ with a two-stage collection approach. Prior implementation of BHQ has used the initial screening questions as the primary method to count and classify events. This one-stage approach

differs from that implemented on the NCVS, which follows up initial victimization screening items with questions that collect more details on what happened. This information is used to classify the event . The second phase compares the use of the final methodology as implemented on audio computer-assisted self- interviewing (ACASI) to one using computer-assisted telephone interviewing (CATI). The analysis will compare the quality of the estimates between ACASI and CATI, as well as to the ongoing NCVS. Data quality will be assessed for coverage/nonresponse bias (CNB) and measurement error (ME). To assess CNB, the study will compare respondent characteristics to benchmark estimates, as well as conduct level of effort analyses to assess the relationship between response rate and bias. Examples of methods to assess ME will include - 1) analysis of incidence and prevalence rates between designs (construct validity), 2) comparison of survey data to narratives of what happened (internal validity), 3) analysis of responses to vignettes (comprehension of questions), and 4) analysis of re-interviews (reliability). The presentation will provide results of the first phase of the project, and discuss more details on strategies for analysis in the second phase.

Minimizing Total Survey Error in the National Crime Victimization Survey: An Assessment of Methods to Adjust for Recall Bias and Fatigue

Andrew Moore[1], Marcus Berzofsky[1], Lance Couzens[1], Dave Heller[1], Chelsea Burfeind[1]
RTI International[1]

This paper follows-up on our 2014 presentation assessing recall bias and fatigue in the National Crime Victimization Survey (NCVS) and will present final results and recommendations. The NCVS is a nationally representative multi-stage household survey conducted by the U.S. Census Bureau for the Bureau of Justice Statistics (BJS) aimed at collecting detailed information about the victims and consequences of crime. The NCVS is designed to allow estimation of annual counts and rates of personal and household victimization by demographic characteristics and to permit comparisons over time. The rotating panel design consists of samples of approximately 50,000 households which are interviewed at six-month intervals over a three year period for a total of seven interviews. All residents 12 years and older in a selected household are interviewed each wave. The NCVS interview consists of a screener in which all respondents are asked about potential types of crime they may have experienced during the prior 6 months and a detailed incident report administered to those indicating a crime during the screener. Given the complex nature of the design, the NCVS is subject to multiple potential sources of error. For example, crime victimization can be highly subject to errors in recalling events, including when the victimization occurred, and/or social desirability bias. Prior to 2006, the first interview served as a bounding interview and was excluded from estimates and the annual data release in an effort to control for respondent telescoping. However, beginning in 2006, to help maintain precision, households that were new to the sample began having their first interview included. Currently, a ratio adjustment is applied to the victimization weights of persons and/or households reporting a victimization during their first interview to correct for potential telescoping. However, this adjustment may be influenced by other survey error sources such as mode effect (since the first interview and later interviews are usually administered via different modes) and fatigue (e.g., there is sizeable panel attrition across the seven interviews especially among younger respondents). In addition, the current approach does not account for interviews that are unbounded due to nonresponse in previous waves or for persons and/or households that enter the sample after the first interview. To account for these additional sources of error, multiple methods have been developed and evaluated. These methods vary not only the type of telescoping adjustment (e.g., overall ratio adjustment, class specific adjustment, model based adjustment) but also who should receive the adjustment (i.e., only first interview respondents or any unbounded respondent) and what group of respondents should comprise the reference group to determine the magnitude of an adjustment for recall bias. The development and incorporation of the fatigue adjustment is also varied and crossed with the different methods for telescoping adjustments. This paper presents an assessment of these methods under a total survey error framework along with results and recommendations for implementation.

Coverage Error in Practice: Operational Issues for Measuring Frame Coverage of Demographic Surveys

Session Chair: Cha-Chi Fan

Is the Whole Sampling Frame Less than the Sum of Its Parts? How Operational Weaknesses Can Undermine Frame Coverage

Clifford Loudermilk[1]
U.S. Census Bureau[1]

Many demographic surveys rely upon traditional household-based sampling frames to provide coverage of their target populations. Typical "building blocks" of household-based sampling frames include address lists, administrative records, and field listings. While each can be very effective in providing coverage of specific subpopulations, there are often

considerable operational challenges in putting these "building blocks" together to create a coherent, effective sampling frame for the entire target population. In addition, frame coverage error can be worsened by failures to put appropriate systems in place to maintain the frame and monitor its performance. We will discuss how operational failures can undermine frame coverage in complex and unexpected ways. Topics of discussion will include "edge effects", operational lags, and the use of address filters.

Assessing Lister Error Associated with Frame Creation - Comparing Address Listing Results and 2010 Census Outcomes

Aliza Kwiat[1]
U. S. Census Bureau[1]

Block Canvassing is one method used to improve survey frames. In theory, block canvassing gives a complete frame because listers see what is on the ground. In practice, listers encounter many challenges in the field that can cause problems with the frame and further impact the bias and variance of survey estimates. This study compared adds and deletes from the 2007 and 2008 Demographic Area Address Listings (DAAL) to 2010 Census outcomes. We looked at the 194 thousand records that were listed by DAAL in 2007 and 2008 and looked at the same records after the 2010 Census.

Selection of Predictors to Model Coverage Errors in the Master Address File

Andrew Raim[1]
US Census Bureau[1]

The U.S. Census Bureau has considered statistical models to help characterize and predict errors on the Master Address File (MAF). This work follows on to Young, Raim & Johnson (Submitted, 2015) and further investigates zero-inflated negative binomial regression to model adds from the 2010 address canvassing operation. We consider several supplemental data sources including the Planning Database, the Longitudinal Employer-Household Dynamics data, and land use data, in addition to the database with outcomes from the operation. Collection of data for address canvassing was subject to a variety of influences not captured in the data. Influences included variations in field representative behavior, in-office post-processing of field data, and other operational details not available at the time of data analysis. Therefore, it is not obvious which predictors explain outcomes from the operation, and variable selection is especially important for this analysis. We carry out an exhaustive variable selection consisting of forward and backward selection steps, and compare candidate models by several likelihood and prediction-based criteria. In contrast to the screening selection used by Young, Raim & Johnson, this method allows us to consider two-way interactions and to rank predictors by their contribution to the model. Residual analysis shows that the obtained model fits well to a majority of the blocks, but the relatively small proportion of blocks which do not fit well tend to be those with the most observed adds. Future work could continue the search for useful predictors and extend the model to support extra variability observed in the data.

A Sensitivity Analysis of Coverage Error for Demographic Surveys

Cha-Chi Fan[1]
US Census Bureau[1]

The main objective of this study is to provide the frame coverage sensitivity profiles for key demographic survey estimates, so that we understand the current quality of the estimates given the existence of coverage error and how changes in living accommodation coverage would impact the quality of the estimates.
Many major demographic surveys sample from a living accommodation frame. Specifically, the Census Master Address File (MAF) was adopted as the primary sampling frame during the 2010 sample redesign. Although the MAF is considered to be the best inventory of the U.S. addresses, maintaining a complete inventory throughout the decade is an extremely difficult task. Recognizing that there is coverage deficiency in the frame, it is important for the Survey Administrators to understand how coverage deficiency in the MAF-based frame would influence the quality of the key survey estimates. Coverage deficiency on the MAF is not necessarily random. Certain types of living accommodations and particular regions of the country are found to have larger coverage concerns. In addition, we observe a spatial relationship in omissions. There is a risk of bias in the

estimates if the information to be estimated is correlated to coverage deficiency. In order to provide quantitative information to help with demographic survey operational decisions, we conduct a sensitivity analysis with a simulation experiment under different coverage scenarios for key survey estimates and establish the frame coverage sensitivity profiles at a national and a state level.

Record Linkage

Session Chair: Frauke Kreuter

Invited Presentation: **Errors in Linking Survey and Administrative Data**

Joe Sakshaug[1]
University of Manchester

Survey and administrative record linkage has become an important tool for increasing research opportunities in the social sciences and is likely to become even more important in the "big data" era as researchers exploit the growing number of data opportunities available to them. While much attention has been given to the many substantive research opportunities that record linkage affords, there has been significantly less attention given to the quality issues associated with linkage. This is an important knowledge gap because the quality of the underlying methods used to link multiple data sources can potentially impact the quality of the inferences obtained from the linked data. This synthesis presents a comprehensive overview of the possible errors that can occur when linking survey and administrative data. We group the possible linkage errors into at least three classes. The first class of errors is due to erroneous linkage. This occurs when a survey record is mistakenly linked to an administrative record that does not belong to the corresponding survey unit. This situation may occur, for example, if a faulty unique identifier (e.g., Social Security number) is reported by the respondent or recorded by the interviewer. Erroneous linkage may also occur when administrative records are defined in terms of one's status in the household (e.g., head of household) and someone other than the target person is interviewed and their responses linked. The second class of errors is caused by imprecise matching variables. When a unique identifier is not available, probabilistic matching is performed on the basis of ambiguous and error-prone identifiers, including name, sex, date of birth, and address. String-comparator metrics and blocking are often used to compute the degree of similarity between two records over all identifying information. However, inconsistencies between the information collected from respondents and the information contained in the administrative database can reduce the quality of the link. The decision rules used to classify pairs of records into links, potential links, and non-links can also impact the quality of the link depending on which matching thresholds are chosen. The third class of errors can occur when respondents do not consent to link their survey responses to administrative data. Informed consent is usually needed to ensure that respondents are aware of the risks and benefits involved in releasing their information for research purposes. Studies have shown that consent rates vary widely from study-to-study and across many disciplines with percentages ranging from the mid-20's to the high 80's. Not only does linkage non-consent reduce the size of the linked database, it can also induce bias in both the survey and administrative variables if differences exist between consenters and non-consenters. Research has shown that age, sex, education, income, and health status, as well as the presentation of the linkage request, are correlated with linkage consent. We demonstrate how to evaluate each of these error sources and assess their impact on the resulting inferences obtained from linked survey and administrative data using real-world examples from linkage projects in the U.S. and in Europe. The paper concludes with a discussion towards the future of survey record linkage and offers general suggestions on how to mitigate linkage errors with the aim of improving the quality of linked data.

The Nature of the Bias when Studying Only Linkable Person Records: Evidence from the American Community Survey

Adela Luque[1]*, Amy O'Hara*[1]*, David Brown*[1]*, Brittany Bond*[2]
U.S. Census Bureau[1]*, U.S. Department of Commerce*[2]

Record linkage across survey and administrative record sources can greatly enrich data and improve their quality while reducing respondent burden and nonresponse follow-up costs. Record linkage can also create statistical bias though. The U.S. Census Bureau makes person records anonymous and linkable across data sources by assigning each record a Protected Identification Key (PIK). However, it is not possible to reliably assign a PIK to every record. Potential non-randomness in PIK assignment can inject bias into statistics using linked data. This paper studies the nature of this bias using the 2009 and 2010 American Community Survey (ACS).

Determining Recall Errors in Retrospective Life Course Data – An Approach Using Linked Survey and Administrative Data

Stefanie Unger[1], Britta Matthes[1]
Institute for Employment Research[1]

Autobiographical recall is imperfect and memories differ in characteristic ways from reality. Therefore, retrospectively reported event histories are at risk of containing a specific memory related error component (recall error). Literature shows that episodes and transitions are underreported, and memory is selectively worse for shorter and atypical events as well as for events that date back a long time. Because of the imperfections in recall, there is often skepticism about the usefulness of retrospective data. One way to analyze retrospective recall errors is using linked survey and administrative data of the same individuals. We can draw on a dataset that contains information on all German employees only excluding self-employed persons and civil servants and link this to our survey data However, the reliability of administrative data has been distrusted, too. It is well-known that its quality depends on the importance of the recorded information for administrative purposes. For instance, information on occupation in the IAB employment data originates from the notification process of the German social security system. The information is unverified (except for simple value checks), and misreporting has no consequences concerning obligations or claims out of the social security neither for the employer nor for the employee. Until now, the recall error of retrospective life course data has not been analyzed while simultaneously considering errors in administrative data. In our presentation we determine the recall errors in retrospective life course data by using the linked data-set ALWA-ADIAB, which combines interview data and administrative data from the same individuals. Assuming that the recall error does not play a role up to three month prior to the interview date, we analyze the "notification error" – the error which can be ascribed to the administrative notification process and by taking this into account, we determine the recall error in the retrospective reports of survey respondents. Using the example of information on occupation we test two hypotheses: First, we assume that the recall of information on occupation is increasingly subject of error when the retrospective interval gets longer. Second, we assume that the quality of survey data is influenced by the stability of job histories. That is, frequent job changes and short times of employment episodes coincide with more recall errors. Our analyses show that the extent of the "notification error" is large and therefore has to be considered when analyzing linked survey and administrative data. In gaining information on the likelihood of error in administrative data as a function of job and firm characteristics, the recall error in survey data can be modeled. This also sheds light on the size and determinants of the recall error when reporting autobiographical events and transitions retrospectively. Both data sources, administrative data and retrospective life course data, have more or less pronounced advantages and disadvantages. The challenge in analyzing such linked data sets is to combine the most reliable information of both sources. Considering the already relatively extensive literature about survey errors this calls for expanding the further research on the reliability of administrative data.

Statistical Analysis of Files Created Through Record Linkage: Joint Modeling of Linkage and Analysis

Michael Larsen[1]
The George Washington University[1]

Record linkage involves bringing together information from two or more files such that the combined records associate together all the available data individual by individual. The product of record linkage is a file with one record per individual that contains all the information about the individual from the multiple files. The problem is difficult when a unique identification key is not available, there are errors in some variables, some data are missing, and files are large. Probabilistic record linkage computes a probability that records from on different files pertain to a single individual or to different people. Some true links are given low probabilities of matching, whereas some non links are given high probabilities. Errors in linkage designations can cause bias in analyses based on the composite data base. The resulting linkage product, in addition to analytical variables, also can contain information on the quality of the linkage and estimated probabilities of correct linkage. Further, information can be made available about record pairs judged not to be correct linkages. Full probability models are proposed for jointly modeling the record linkage process and subsequent statistical analysis.

Use of Linked Survey Data to Develop Responsive Design Sampling Strategies in the Medical Expenditure Panel Survey

Lisa Mirel[1], Sadeq Chowdhury[1], Steven Machlin[1]
DHHS\AHRQ[1]

Using information from a prior survey that is linked to a new survey can help develop responsive design sampling strategies. One example of this is the linkage between the National Health Interview Survey (NHIS) and the Medical

Expenditure Panel Survey (MEPS). MEPS is a complex, multi-stage, nationally representative sample of the U.S. civilian noninstitutionalized population. Each year a new sample is drawn as a subsample of households from the prior year's NHIS. Whether the NHIS interview was complete or partially complete is associated with MEPS response propensity and is currently being used as a sampling stratum in MEPS. This paper describes an evaluation to use other NHIS variables (e.g. the interviewers' assessment of the likelihood of response in a linked survey) and certain MEPS paradata variables to increase response rates and reduce data collection efforts. We will describe the utility of the NHIS and MEPS paradata variables for sampling and their potential impact on variance. Lastly, we will discuss plans for future work in this area.

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

Joe Sakshaug[1], Brady West[1]
University of Manchester, Institute for Social Research

The survey methodology literature is replete with alternative descriptions of the Total Survey Error (TSE) paradigm. 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, but 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 design-based or model-based estimation methods that correctly account for important 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 apparent analytic errors (given that word limits may prevent a researcher from fully describing what was done in a given analysis) by performing a meta-analysis of 100 sampled publications 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. We find that several types of apparent analytic errors are quite prevalent, including inappropriate subpopulation analyses and a failure to use appropriate software. Analysts also seemingly 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 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.

Sunday, September 20, 2015
1:30 - 3:30 p.m.
Paper Session II

Errors in Panel Surveys
Session Chair: Brad Edwards

Invited Presentation: **Total Survey Error for Longitudinal Surveys**

Peter Lynn[1], Peter J. Lugtig[1]
ISER[1]

The aim of this paper is to describe the application of the total survey error paradigm to longitudinal surveys, by which we mean surveys that collect data on multiple occasions from the same sample elements. Longitudinal surveys are of great importance and considerable investment has been made in them by government agencies and academic funding bodies in recent years. There are several aspects of survey error, and of the interactions between different types of error, that are distinct in the longitudinal survey context. Furthermore, error trade-off decisions in survey design and implementation are subject to some unique considerations. For these reasons, a framework for total survey error in the longitudinal survey context is desirable. We aim to provide such a framework, dealing with conceptualisation of errors, study of interactions between errors, and decision making in the presence of trade-offs. The paper will begin by briefly introducing the concept and nature of longitudinal surveys, as relevant to total survey error. We will here present the case for longitudinal surveys needing special consideration, pointing out that the existing literature on total survey error is largely restricted to cross-sectional surveys, at least implicitly. We assume that we will not need to summarise current thinking and practice regarding TSE generally and will instead be able to refer to an introductory chapter in the volume. The next section will provide an overview of error sources in longitudinal surveys, focussing on those aspects that are distinct from the cross-sectional survey case. For example, we will point out that key estimates are typically measures of change or stability over waves, and therefore that the correlation of measurement errors over waves is often more important than the nature of error in the measure collected at any particular wave. Similarly, we will explain how non-response error is typically the cumulative result of initial non-response and subsequent attrition. We will then discuss how errors from different sources can interact on longitudinal surveys and how these interactions can change over the life of the survey, as more waves of data collection are carried out. We will outline ways in which such interactions can be studied and how the information gained from these studies can be used to inform design and implementation decisions. We will draw a distinction between surveys with a limited number of waves that is known from the outset and those that are indefinite (where the ultimate number of waves typically depends on continued success at raising funds). We will illustrate our discussion of interactions and trade-offs between errors with a small number of examples from real surveys. We will aim to include examples from different types of longitudinal surveys (e.g. household panels, individual cohorts) and different data collection modes (e.g. CAPI, web). The paper will end with a discussion section, in which we will draw attention to the interplay between TSE, other dimensions of survey quality (e.g. timeliness, relevance), and survey costs. We will outline some of the ways in which this interplay may manifest itself in the longitudinal survey context (but will not discuss the details in much depth). In this section we will also highlight areas of research or practice that we believe deserve more attention regarding TSE for longitudinal surveys.

Invited Presentation: Using Doorstep Concerns Data to Evaluate and Correct for Nonresponse Error Components in a Longitudinal Survey

Ting Yan[1], Shirley Tsai[1]
Westat[1]

Doorstep concerns – one type of paradata – capture the interactions between interviewers and potential survey respondents during the survey introduction and reveal the concerns sampled members have expressed about the survey request and also their reasons for refusing the survey request when refusal occurs. We've created two parsimonious measures that retain the interrelationships inherent in the doorstep concerns data – the Perceived Concerns Index (PCI) created through principal component analysis and the Reluctance Class (RC) generated by latent class analysis – and have demonstrated that both measures are effective in characterizing and assessing the level of reluctance exhibited by potential survey respondents (Yan and Tsai, 2012; 2013). Extending the earlier research, this paper explores the use of the two summary measures in a longitudinal survey setting to understand the relationships between different error components under the Total Survey Error framework. Specifically, we focus on two kinds of nonresponse error – unit nonresponse error and item nonresponse error. According to the response continuum model posited in Yan and Curtin (2010), there is a positive relation between a person's likelihood to participate in a survey and his/her likelihood to answer survey questions. This positive relationship translates into a trade-off between the two types of nonresponse errors – efforts to reduce unit nonresponse could increase item nonresponse. This trade-off becomes critical for longitudinal surveys where respondents are invited back periodically to participate in more rounds of interview. It is necessary to study the trade-off between unit and item nonresponse biases and to track the changes in the total nonresponse error as respondents are being asked to participate in later waves of interview. Taking advantage of data from four waves of the Consumer Expenditure Interview Survey (CE), we will conduct three analyses. Analysis 1 tests the response continuum model by examining the relationship between a person's reluctance (as captured in the doorstep concerns data) and their likelihood to participate in and to provide missing data in later waves of the CE. Using wave 2 data, Analysis 2 will empirically examine bias due to unit nonresponse and bias due to item nonresponse at later waves of CE (i.e., at waves 3, 4, and 5) and track the changes in total nonresponse bias by wave. Of particular interest, we will compare both the component nonresponse biases and total nonresponse bias by levels of reluctance exhibited by respondents as captured in

the doorstep concerns data. Analysis 3 will explore the use of the two measures of reluctance in nonresponse adjustment. We plan to introduce PCI and RC in creating nonresponse adjustment cells and examine the changes in estimates of key variables of interest before and after including PCI and RC. We conclude this paper by discussing the potential use of PCI and RC in an adaptive design in a longitudinal survey setting to make informed decisions on attempts to reduce total nonresponse bias in later waves of data collection.

Mobile Device Surveys in a U.S. College Student Population: Results from a Program of Research Exploring Nonresponse and Data Quality Issues in a Longitudinal Panel Survey.

Scott Beach[1], Donald Musa[1], Stephen Strotmeyer[1], Janet Schlarb[1]
University of Pittsburgh[1]

Increasing numbers of surveys worldwide are being completed on mobile devices like Smartphones. One question is whether the quality of data collected with mobile technologies differs from data collected via more traditional methods like landline phones or laptop / personal computers. Studies comparing telephone survey data collected by interviewers on cell phones versus landlines have shown minimal differences in data quality, but less work has been done comparing self-administered web survey data collected via mobile devices versus more traditional personal computers/laptops. Web surveys completed on mobile devices may be of lower quality due to factors such as readability, distraction, multi-tasking, and the presence of others in social situations, which may lead to survey breakoffs or higher levels of measurement error. These issues are particularly relevant to surveys of college students, who are more likely than older adults to use Smartphones for email / internet activity. This paper examines data collected from randomly selected college students from the University of Pittsburgh (U.S.) in 2013 (n=1,626), 2014 (n=1,431), and 2015 (data still being collected) as part of a longitudinal panel survey of undergraduate students. Respondents are categorized as mobile (Smartphone, Android, etc.) or non-mobile (PC, laptop, Ipad, etc.) based on the device used to take the survey. One part of the paper focuses on the impact of several survey design changes implemented in 2015 in response to recent declines in overall response rates, including (1) optimization of survey formatting for device type (no grids, 1-3 questions per page with scrolling on mobile vs. grid questions, minimal scrolling on non-mobile); (2) overall shortening of survey length from 95 to 60 questions; (3) an explicit statement in the email invitation that that the survey can be taken with and is formatted for either device type; and (4) official announcements about the survey on the University student resource portal website. Outcomes include overall response rates and usage of mobile devices across years; and comparisons within and across years of mobile versus non-mobile respondents on breakoff rates, item missing data, survey completion times, quality of open-ended responses, and inter-correlations among specific and general student satisfaction indicators. Preliminary 2015 results show both an increased overall response rate and more usage of mobile devices compared to 2013-14. We also present results from a 2015 randomized experiment with students not in the longitudinal panel (approximate base n=2,000) varying (1) inclusion vs. exclusion of an explicit statement in the email invitations that the survey can be taken on either device type, crossed with (2) formatting for mobile vs. non-mobile device as described above, regardless of device type used (i.e., mobile formatting for mobile device; non-mobile formatting for mobile device; non-mobile formatting for non-mobile device; mobile formatting for non-mobile device). Respondents will be randomly assigned to survey format once they are categorized by device type after beginning the survey. The nonresponse and data quality outcomes noted above will be examined as dependent variables. Theoretical and practical implications of results of both the observational and experimental studies will be discussed.

Introducing Adaptive Design Elements in the Panel Study "Labour Market and Social Security" (PASS)

Mark Trappmann[1], Gerrit Mueller[2]
IAB University of Bamberg[1], IAB[2]

PASS is one of the major German panel surveys. It focuses on unemployment and poverty dynamics. Since 2007 about 15.000 persons in about 10.000 households are interviewed each year. PASS uses a sequential mixed-mode design of CAPI and CATI. Data can be linked to detailed administrative records on employment histories for all respondents who provide informed consent. Since Wave 4 detailed paradata have been available on a biweekly basis during fieldwork. Since Wave 6 (2012) these have been used for informed interventions into the fieldwork of the panel. The presentation gives an overview of the fieldwork monitoring based on paradata in combination with frame data and the adaptive survey design. The adaptive survey design comprises experiments concerning optimal contact times, wording of advance letters and cash incentives for temporary dropouts and prioritizing low propensity cases in later phases of the fieldwork. In the presentation a focus will be on a series of interventions aiming at prioritizing low propensity sample members. For these experiments response propensities were estimated for CAPI cases during fieldwork based on contact histories and frame data. In the last phase of data collection of wave 7 interviewers were promised considerable premiums for completing cases with a low predicted response propensity. The premium was offered for a random half of the low propensity cases. In

wave 8 the experiment was repeated crossing interviewer incentives with additional prepaid respondent incentives. We find that interviewer incentives lead to a higher probability of receiving a final status (interview or refusal) while the number of cases still open at the end of the fieldwork (address problems, noncontacts, broken appointments) decreases. However, response rates are only significantly higher for the experimental group when they are complemented by incentvies on the respondent side. In addition, high quality frame data enable us to assess the impact of the adaptive design on nonresponse bias.

Interviewer Effects Following Throughout the Total Survey Error Framework

Session chair: Annelies Blom

Interviewer Effects on Multiple Sources of Survey Error

Daniela Ackermann-Piek[1], Annelies G. Blom[1]
University of Mannheim[1] *infrastructure*

Concerns about interviewer effects in interviewer-mediated surveys have accompanied generations of survey researchers. As early as the late 1920s, Rice (1929) found that interviewers introduce measurement bias. However, interviewers' influence is not limited to measurement error, but affects nearly all aspects of survey errors, including sampling, nonresponse, coding and editing of survey responses. Following the Total Survey Error framework, our paper shows how interviewers can cause errors in multiple areas of a survey. We focus on interviewer effects on measurement error and interviewer effects on nonresponse. For this we draw on various interlinked data sources collected during the first wave of the German implementation of the Programme for the International Assessment of Adult Competencies (PIAAC). The analyses into measurement error and nonresponse are supplemented with interviewer characteristics and attitudes collected during an interviewer survey similar to the one described by Blom and Korbmacher (2013). The analyses build on own prior research into interviewer effects on the individual error sources. Using PIAAC data, we are in the fortunate situation of being able to combine the results of analyses into individual error sources and look into the joint interviewer effect on survey errors. Literature Blom, A. G., & Korbmacher, J. M. (2013). Measuring interviewer characteristics pertinent to social surveys: A conceptual framework. Survey Methods: Insights from the Field, 1(1). doi: 10.13094/SMIF-2013-00001 Rice, S. (1929). Contagious bias in the interview: A methodological note. American Journal of Sociology, 35(3), 420-423.

Interviewer Effects on Straight-lining

Geert Loosveldt[1], Koen Beullens[1]
KULeuven[1]

Many survey questionnaires contain lists of objects and/or lists of statements about a particular topic with the same response categories. Researchers assume that respondents can and will differentiate the objects or items on the list and that they use different response options to differentiate their ratings. However researchers are also aware that respondents sometimes provide the same answers to all questions in a block of questions about the same topic or object. This response style is called non-differentiation or straight-lining. Research about different types of non-differentiation (ARS, DRS, MRS and ERS) conclude that these response styles are stable individual characteristics (Weijters, Geuens & Schillewaert, 2010). This means that the respondent is the main responsible for straight-lining and that respondent's characteristics and personality are relevant to explain response styles in general and straight-lining in particular. However demographic and personality variables explain only a relatively small proportion of the variance of response styles. On the other hand culture and country-level characteristics seem to explain a relatively large proportion of the variance of response styles (Van Vaerenbergh and Thomas, 2013). This result indicates that not only respondents are responsible for their response style but response styles can also be affected by contextual factors. One of the contextual or situational factors discussed in the paper is the presence of interviewers in face-to-face interviews and their impact on straight-lining. Although the factor 'interviewer' is not completely absent in the research about response styles in face- to-face interviews, interviewers certainly do not play a dominant role in this type of research. Usually interviewers and respondents are considered as two separated sources of measurement error. In this paper we assume that the impact of both sources are not completely independent and that measurement error due to response style in not just a matter of the respondent's cognitive efforts but is also affected by the way interviewers are dealing with this particular response behavior. This means that interviewers are partially responsible for the amount of straight-lining they obtain and we expect that interviewers explain a significant proportion of the variance in straight-lining. ESS data of round six will be used to test our assumptions and expectations.

Do Interviewers with High Cooperation Rates Behave Differently? Interviewer Cooperation Rates and Interview Behaviors

Kristen Olson[1], Jolene Smyth[1], Antje Kirchner[1]
University of Nebraska-Lincoln[1]

Interviewer skills for obtaining cooperation from sampled households require flexibility, tailoring to the respondent, and maintaining interaction (Groves and Couper 1998). Furthermore, the most successful interviewers at gaining response rates often deviate the most from scripted introductions (Snijkers, et al. 1999; Houtkoop-Steenstra and van den Bergh, 2000). On the other hand, administration of survey questionnaires requires standardized adherence to best practices, reading questions exactly as written, nondirective probes, and a clear set of regimented behaviors during the interview (Fowler and Mangione 1990). That is, interviewers are required to be flexible during recruitment, but standardized during measurement. These skill sets may be at odds. Recent research (Brunton-Smith et al., 2012) has shown a U-shaped relationship between interviewer cooperation rates and interviewer variance: The least and the most successful interviewers have the largest interviewer variance component. Still open is why this association occurs. One hypothesis is that interviewers with higher cooperation rates act differently during a survey interview than other interviewers, translating the flexible style from recruitment into survey administration. This paper will examine behavioral differences in the survey interview between interviewers who are more successful at gaining cooperation and those who are less successful. We build off of previous research (Kirchner and Olson 2014) that showed that interviewers with higher cooperation rates have shorter interviews and more answer changes. We now examine whether interviewer behaviors differ for interviewers with higher versus lower cooperation rates. In particular, we examine question misreadings, probing, feedback, disfluencies, and clarifications. We hypothesize that interviewers with higher cooperation rates will have more rapport behaviors and fewer behaviors such as probing or clarification. We use the Work and Leisure Today Survey (n=450, AAPOR RR1=4.7%), including survey data, paradata, and behavior codes. Preliminary analyses indicate that interviewers with higher cooperation rates deviate more from the question wording, introducing (major) changes to the question stem or response options more often than interviewers with lower cooperation rates. They also tend to not repeat the respondent's answer appropriately when verifying a given response. Interviewers with lower cooperation rates, on the other hand, show more disfluencies when reading a question and do not laugh as often. The paper will conclude with implications for interviewer training and questionnaire design.

Improving Survey Data Quality Under Total Survey Error Framework: Application to Two Surveys of U.S. NSF – National Survey College Graduates (NSCG) and Survey of Doctorate Recipients (SDR)

Session Chair: Donsig Jang

The Effects of Data Editing and Imputation on Total Survey Error of the National Survey of College Graduates

Alicia Haelen[1], Donsig Jang[1]
Mathematica Policy Research[1]

Survey data are prone to substantial bias and variance during each phase of survey research, including instrument design, data collection, processing, and estimation. This presentation looks at the impacts of each source of survey error, with a keen focus on data processing using the National Survey of College Graduates (NSCG). Although survey contractors devote considerable time and effort to assessing the impact of data editing and imputation, little work has been done to evaluate the editing and imputation procedures for this survey. Consequentially, we have minimal knowledge of the extent to which the editing and imputation procedures influence final survey estimates Our current research effort is a comprehensive attempt to fill this knowledge gap and gain a better understanding of the existing editing and imputation processes, their impact on survey estimates and total survey error, and develop alternative methodology that improves precision and efficiency. We present findings from a simulation study conducted on different editing and imputation procedures, comparing treatment rates and key survey estimates, and offer recommendations for improving the NSCG procedures. The expectation is that the research results will be available for possible implementation as part of the 2015 survey cycle.

Improving Data Quality During Data Collection: Adaptive Design in the 2015 NSCG

Stephanie Coffey[1], Benjamin Reist[2]
U.S. Census Bureau / Joint Program in Survey Methodology[1], U.S. Census Bureau[2]

In 2013, the National Survey of College Graduates (NSCG) included an adaptive design methodology study to explore whether data monitoring techniques and data collection interventions could be implemented during the data collection effort of a large-scale national survey. Results of this study showed that it is possible to monitor data as it is being collected and use the monitoring to inform data collection interventions that have potential to improve data quality and reduce survey costs. Given our operational success with incorporating adaptive design functionality into a large-scale national survey, we included a follow-up adaptive design study in the 2015 NSCG. This new study included a larger sample design, increasing the power of statistical comparisons between the study and the control group on a variety of metrics, including response rate, R-indicators, cost, and effect on key estimates. In addition, the 2015 NSCG methodology study took advantage of the longitudinal nature of the NSCG data, and explored the use of adaptive design techniques on both new and returning sample members. At the same time, however, data collection interventions aimed at improving these measures (and thereby addressing issues most important for the NSCG like nonresponse bias and data timeliness) potentially cause error tradeoffs between response and measurement. This talk presents the effects of the 2013 adaptive design study within a total survey error framework, and attempts to answer several questions: Did the respondents in the experimental groups have distributions of frame variables more similar to the full population than the control group? Were distributions of response mode different between the experimental groups and the control group? Are the distributions of key estimates different in the experimental groups versus the control group? Can we attempt to validate response quality externally? In order for adaptive design methodology to advance, studies that result in definitive statements about the impact of various data monitoring techniques and data collection interventions must be conducted and documented. The hope is that the 2015 NSCG study will add significantly to the existing survey methodology literature on adaptive design.

Investigating Non-sampling Error in Longitudinal Panel of the Survey of Doctorate Recipients

Wan-Ying Chang[1], Lynn Milan[2]
National Science Foundation / NCSES[1], National Science Foundation / NCSES[2]

The Survey of Doctorate Recipients (SDR) is a biennial survey conducted since 1973 on individuals with a U.S. research doctoral degree in a science, engineering, or health field. It used a design that carries forward a longitudinal panel from previous survey cycles and refreshes it in the new survey cycle by adding a sample of recent doctoral degree earners. Maintaining the panel not only provides a longitudinal profile of individuals but also significantly reduces the cost of data collection by decreasing the number of sample members who need to be located. As the panel size grows over time, it it necessary to define efficient trimming strategies in order to control the overall sample size and survey cost. Various design options for trimming the panel based on participation history have been proposed in the past; however, little is known about the association between response pattern and different types of non-sampling error. This study used data from the most recent four SDR survey rounds to investigate the association between response pattern and data quality in terms of sample representativeness, non-response bias, and measurement error. The finds will be used to evaluate different design options for maintaining the panel under the total survey error framework.

The TSE Framework: Looking Back and Looking Forwards

Session Chair: Paul Biemer

Invited Presentation: **The Roots of the Concept of Total Survey Error**

Lars Lyberg[1], Diana Stukel[2]
Stockholm University[1], FHI360[2]

Applying the Total Survey Error Paradigm to Multiple Surveys and Auxiliary Data

Tom Smith[1]
NORC[1]

Tom W. Smith The total survey error (TSE) paradigm was originally developed to apply to single surveys. As such it is both a theoretically and an empirically powerful approach to identifying and reducing survey error and thereby increasing the reliability and validity. But surveys are often not used in isolation. Surveys are usually most valuable when they are combined together, such as in a) comparative research (e.g. cross-national and cross-cultural), b) time-series or trend analysis, and c) longitudinal or panel analysis. In each of these designs, multiple surveys are utilized in the research designs. TSE can fruitfully be used not only to improve each of the individual surveys in such multi-survey designs, but also to maximize comparability across surveys (Smith, 2011). For example, in comparative research the goal is to achieve functional equivalence in surveys across nations and cultures. This goal can be advanced by using the TSE approach in

general and by the introduction of the TSE concept of comparability error in particular. Just as research and analysis are enhanced when multiple surveys are utilized, research can be advanced and findings made more robust when surveys are augmented with other auxiliary data from other sources. For example, the multi-level, multi-source (MLMS) approach (Smith and Kim, 2013a; 2013b) shows how both individual- and aggregate-level data from auxiliary sources such as Censuses, GPS databases, and administrative records can be used to augment surveys. Of course, the utilization of multiple data sources necessarily makes error structures more complex and that means that more attention must be paid to understanding and minimizing error. Moreover, this introduces both new sources of error (e.g. linkage error from the individual- and aggregate-level merging of data from auxiliary sources to surveys) and the likelihood that error interactions and correlations will occur (e.g. that survey cases with less complete information will have more linkage error or that respondents who provide less accurate information on surveys will also have less accurate data about themselves in auxiliary sources). Using TSE will be very important initially in identifying and organizing these additional sources of error and ultimately in modelling and minimizing all error. Besides integrating auxiliary data with survey data at both the individual- and aggregate-level, auxiliary sources can be used to complement survey data at the analysis stage. As Otis Dudley Duncan's "outrigger principle" indicates (Turner and Martin, 1985), survey results are made stronger (e.g. more reliable and valid) when they are supplemented by results from external sources (both survey and non-survey). For example, studying trends in household gun ownership can be improved when results from time-series surveys are combined with data from the Census on household size and composition, the issuance of hunting licenses, figures on the production and import/export of firearms, the volume of background checks to purchase firearms, the Uniform Crime Reports of victimizations, etc. And once again, TSE should play a central role in combining together these various sources and understanding how they can best complement the survey results.

Adapting and Applying the TSE Paradigm to All Quantitative and Qualitative Research

Paul Lavrakas[1]
Private Consulting[1]

The Total Survey Error (TSE) paradigm, and its virtues, has been articulated by various scholars during the last half century, but none more importantly or extensively than Groves in his seminal volume, Survey Errors and Survey Costs (Groves, 1989). These explanations have focused on advancing a "way of thinking about" survey research framework – the TSE – meant primarily to aid survey researchers in their evaluation of the reliability and validity of sample surveys and/or to help them conceptualize more reliable and valid surveys of their own. The vast majority of TSE-related scholarly work has focused on the development and application of statistical and methodology methods that investigate specific aspects of TSE within particular research studies. This, for example, includes work investigating within-unit coverage errors, nonresponse bias, adjustment errors, respondent-related measurement errors, as well as other components of the TSE paradigm. In contrast, there is relatively little written that addresses the ways in which the TSE paradigm can be adapted and applied in novel ways beyond the domain of survey research. An exception is the work of Lavrakas (2013) who explicitly called attention to this in his Presidential Address to the American Association for Public Opinion Research. Because the TSE paradigm is a comprehensive framework that addresses all important aspects of validity and reliability in social research, Lavrakas has argued that it easily can (and should) be adapted and applied to thinking critically about the planning, implementing, and interpreting of any type of social research. To that end, Lavrakas has recommended that scholars and practitioners who conduct social research other than surveys utilize a Total Focus Group Error framework, a Total Observational Error framework, a Total Content Analysis Error framework, and so on, to guide own their research studies. Thus, Lavrakas has articulated and strongly advocated the use of a "Total Error" (TE) perspective with all forms of research used in the social, behavioral, and marketing sciences. And this call includes applying a Total Error mindset to both quantitative and qualitative research designs. As it applies to qualitative research, Roller and Lavrakas (2015, forthcoming) have adapted the TSE paradigm to create a Total Quality Framework (TQF) using language and concepts with which qualitative researchers are familiar and comfortable. The components of the TQF are Credibility, Analyzability, Transparency, and Usefulness. The TQF maps well to the TSE framework: for example, Coverage Error, Sampling Error, and Nonresponse Error all fit within the "Scope" component of Credibility in the TQF. Therefore, the presentation and chapter proposed via this abstract for consideration for the 2015 International Total Survey Error conference (and its related monograph) will be about the myriad ways that the TSE paradigm can and should be applied across social, behavioral, and marketing science disciplines. It will articulate a generic "Total Error" paradigm for use in quantitative research and a "Total Quality Framework" for use in qualitative research. It will provide many examples of how these perspectives can and should be applied, including when conducting literature reviews, writing RFPs and proposals, and evaluating the quality of research-based legal evidence.

Infrastructure for the Use of Big Data to Understand TSE: Examples From Four Survey Research Organizations

Session Chair: Fritz Scheurenva

Invited Presentation: Big Data Infrastructure at the Institute for Employment Research (IAB)

Antje Kirchner[1], Daniela Hochfellner[1], Stefan Bender[1]
University of Nebraska-Lincoln[1]

This paper outlines the purpose and infrastructure developed to manage and analyze administrative and survey data at the Institute for Employment Research (IAB) of the German Federal Employment Agency (BA). Our focus lies on discussing two possibilities of how administrative and survey data may be used for an assessment of TSE: Data quality in surveys and sampling. We review limitations and assumptions when working with administrative, process-generated data that are related to accuracy, completeness, timeliness and linkage issues by exploiting two datasets that are currently available at IAB. The first is individual level data, the linked panel 'Labor Market and Social Security' and administrative data (PASS ADIAB), and the second dataset is on establishments, the 'Linked Employer-Employee Data (LIAB)'. Furthermore this paper gives insides on how IAB updates and expands its data products to offer new possibilities for improving the quality of surveys and assessing TSE, as well as how data access is provided to the scientific community.

Invited Presentation: Big Data Serving Survey Research: Experiences at the University of Michigan Survey Research Center

Gregg Peterson[1], Grant Benson[1], Frost Hubbard[1]
University of Michigan-Ann Arbor[1]

Big Data offers a promise of significant reduction in sampling and respondent identification costs through the enrichment of sampling frames. However, due to the unstructured nature of many of the data sources, the reliability and coverage property of these data enhancements are often in doubt. In this paper we compare the data appended to traditional sampling frames by two commercial vendors with results obtained by SRC through interviewer-mediated screening.

We find that while the data obtained from Big Data are imperfect, the appended data are more reliable in some domains than in others. We also find that by better understanding the business logic used by data providers to enhance the data, we are able to further improve on the data reliability by requesting minor process changes of the vendors.

Invited Presentation: Using "Big Data" to Evaluate Survey Data: Lessons Learned at the U.S. Census Bureau

Elizaeth Nichols[1], Mary H. Mulry[1], Jennifer Hunter Childs[1]
U.S, Census Bureau[1]

Currently the U.S. Census Bureau is engaging in a large research program aimed at increasing the usage of data from federal and third-party administrative records in the production of its population, housing and business censuses and surveys. The goals of the research are to increase efficiency, reduce cost, improve quality, produce new products, and reduce respondent burden. As part of this research, the U.S. Census Bureau conducted two separate research projects matching survey data to administrative records and third-party data to measure response error in reported move dates. In the first project, we compared the self-reported move date from selected years of the Bureau of Labor Statistics sponsored National Longitudinal Survey of Youth, 1997 (NLSY97) cohort to records in the commercial database Accurint. In the second project, we compared self-reported move dates from a Census Bureau sponsored survey to data from the U.S. Postal Service National Change of Address (NCOA) files for March and April of 2010.

In both projects, we started with large amounts of data in the form of administrative records. In both projects, we found that our comparison sources had their own error structures that presented challenges during analysis. For the Accurint database, we revisited our matching between the survey data and the administrative records and spent much of our time trying to arrive at a "clean" dataset. By "clean," we mean a dataset in which we were confident that the links between the survey and administrative records were for the same person and the same event. Only after considerable effort did the dataset appear suitable for drawing conclusions about the response error in move dates. Similar challenges faced us in the second project using the NCOA database. Ultimately, in each project only a small fraction of the original cases that we had at the beginning of the projects were appropriate for inclusion in our analyses. This paper discusses our experience using the administrative data as a comparison source for survey data and our lessons learned when preparing to combine survey reports with administrative records.

Invited Presentation: Statistics New Zealand's Approach to Making Use of Alernative Data Sources in an Era of Integrated Data

Felipa Zabala[1], Anders Holmberg[1], Christine Bycroft1, Giles Reid[1]
Statistics New Zealand[1]

Statistics New Zealand is progressing a modernisation programme called *Statistics 2020 Te Kāpehu Whetū.* A key part of the programme is increasing the use of indirectly collected data from previously unused sources. This means a change from delivering official statistics that mainly depend on sample surveys, to a situation where the data are often integrated from multiple sources and, in the collection phases, (at least initially) less tailored for the end-purpose.

To explain to stakeholders the usefulness of these statistics (stakeholder utility), as well as to identify flaws in survey design and quality, Statistics New Zealand has applied a comparative framework, which is based on the total survey error approach. In the paper we describe this work and present how various error components have been identified and evaluated. We discuss the lessons learned from investigating new sources of data and how this can be applied to the design considerations of New Zealand's future population census. To modernise a census is one of the biggest redesign tasks a national statistical organisation can undertake. Exploring the option of using alternative data sources to a traditional census collection highlights the importance of exhaustive and transparent methods to assess statistical quality. It also provides a showcase for the use of a total survey error framework in other circumstances such as Big Data applications

Assessing Survey Response Error Using Administrative Records and Evaluating Administrative Records Coverage

Session Chair: Brad Edwards

Invited Presentation: Estimating Error Rates in Administrative Registers Using Latent Variable Modeling: An Application and Validation Study

Daniel Oberski[1]
Tillburg University[1]

Administrative register data are increasingly used worldwide to replace or supplement the census (Wallgren & Wallgren 2007), and are thought to provide a cost-effective opportunity for longitudinal full-population data analysis in the social sciences (Entwisle & Elias 2013). They are also frequently used as "validation data" to study measurement error in survey questions (e.g. Kreuter et al. 2010). In spite of quality control procedures, however, there are strong indications that administrative register data can themselves contain considerable measurement error. Moreover, typically the error process does not conform to classical measurement error models. Such errors negate the potential usefulness of administrative data, making it essential to evaluate their extent. We discuss latent variable modeling as a way to estimate measurement error in administrative data by combining error-prone administrative data with an error-prone survey. To demonstrate the approach, a latent class model is applied to linked register-survey residence data from the municipality of Amsterdam. Moreover, we validate the approach by comparing the estimates obtained through latent variable modeling with estimates of the error rates obtained from audits performed by the municipality. Compared with such audits, latent variable modeling is shown to be a highly cost-effective method of evaluating the extent of measurement error in administrative data.

Coverage of Children in Administrative Records and Census Data

Catherine Massey[1]
U.S. Census Bureau[1]

Demographic analyses using vital records consistently reveal undercounts of children in decennial censuses. The greatest undercounting occurs in the youngest age categories. For the 2010 Census, O'Hare (2012) estimates a 4.6 percent undercount of children ages zero to four. This type of undercounting is not unique to the decennial census, and research has exposed undercounts of children in survey data such as the American Community Survey. Administrative records can potentially ameliorate this problem and enhance our understanding of the systematic undercount of children. This paper documents the coverage of children in administrative records and describes the distribution of record coverage across age groups for multiple federal and state administrative records, as well as commercial data sources. This paper also examines the individual and household characteristics associated with undercounted children. Using linked 2010 Census and

administrative records, I examine the characteristics of children and their families who are in administrative records, but were missing in the 2010 Census. I also document children in the 2010 Census who are missing in administrative data. Regression analysis identifies factors associated with both types of undercount.

Assessing the Effect of the Tipped Minimum Wage Using W-2 Data
Maggie Jones[1]
U.S. Census Bureau[1]

Previous research on tipped wages has been hampered by the reliance on surveyed self-reported earnings. Respondents may have difficulty remembering what they earn in tips over a set period, and tips may vary dramatically both at the shift level and seasonally. However, employers must report tip income separately from wages on W-2s. Using W-2 data linked to the CPS ASEC, I present evidence on the reporting of tips by employers and assess how the information compares with self-reports. I then analyze the effect of the tipped minimum wage on wages, tips, employment, and hours worked for tipped employees in the restaurant industry.

Medicare Coverage and Reporting of the Elderly Population: A Comparison of CPS and Administrative Records
Renuka Bhaskar[1], *James Noon*[1], *Sonya Rastogi*[1], *Brett O'Hara*[1], *Victoria Velkoff*[1]
U.S. Census Bureau[1]

Medicare coverage of the elderly population in the United States is widely recognized as being nearly universal. Recent statistics from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) indicate that 93.1 percent of individuals ages 65 and older were covered by Medicare in 2013. Those without Medicare include those who are not eligible for the public health program, though the CPS ASEC estimate may also be impacted by misreporting. Using linked data from the CPS ASEC and Medicare Enrollment Database, we estimate the extent to which individuals misreport their Medicare coverage - including those who report having Medicare but are not enrolled (false positives) and those that do not report having Medicare but are enrolled (false negatives). We use regression analyses to evaluate factors associated with both types of misreporting including socioeconomic, demographic, and household characteristics. We then provide national and state-level estimates of the implied Medicare-covered, insured, and uninsured elderly population, taking into account misreporting in the CPS ASEC. Finally, we evaluate the characteristics of those who are not covered by Medicare and examine separately those who are uninsured and those who have only private insurance, Medicaid, or SSI. Our results will be useful to researchers studying insurance coverage and reporting and to policy makers aimed at improving health insurance coverage for the elderly population.

Response Error and the Medicaid Undercount in the Current Population Survey
James Noon[1], *Leticia Fernandez*[1], *Sonya Rastogi*[1]
U.S. Census Bureau[1]

The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) is an important source for estimates of the uninsured population. Medicaid coverage is one component of uninsured estimates, and previous research has shown that survey estimates consistently produce an undercount of beneficiaries compared to Medicaid enrollment records. This paper extends past work by examining the Medicaid undercount in the 2007-2009 CPS ASEC as compared to enrollment data from the National Medicaid Statistical Information System (MSIS) for calendar years 2006-2008. Linking individuals across datasets, we analyze two types of response error regarding Medicaid enrollment. First, some persons have Medicaid coverage but report no coverage in the CPS ASEC (false negatives). Second, some persons report Medicaid coverage in the CPS ASEC but can not be linked to Medicaid enrollment data (false positives). We use regression analysis to analyze factors associated with false negatives and false positives in the 2009 CPS ASEC and discuss implications for estimating the uninsured population.

Monday, September 21, 2015
1:30-3:00 p.m.
Paper Session III

Nonresponse and Measurement Error I
Session Chair: Steve Cohen

Nonresponse and Measurement Bias in the American Time Use Survey

John Dixon[1]
BLS[1]

The American Time Use Survey (ATUS) is designed to measure how people spend their time. The ATUS sample is drawn from households completing their final month of interviews for the Current Population Survey (CPS). Because the CPS collects a wealth of demographic information about respondents, this design provides information about ATUS nonrespondents. Measurement error, due to either forgetting an activity or mistakes estimating the time or duration, may also contribute to bias in the ATUS survey. This paper focuses on nonresponse bias and measurement error. A propensity score model is used to examine differences in time-use patterns between those who are likely to respond and those who are reluctant, and to assess the extent of nonresponse bias. The two processes (forgetting and duration recall) will be explored by looking at the large number of zeros in many time-use categories, with the zeros serving as indicators of possible forgetting, and unusual durations conditional on remembering indicating possible recall error.

Did MEPS Round 1 Field Period change designed to improve round 1 response rates inadvertently affect MEPS utilization reporting in 2011

Frances Chevarley[1]*, Karen Davis*[1]
AHRQ[1]

This paper analyzes estimates from the Medical Expenditure Panel Survey (MEPS) matched with the National Health Interview Survey (NHIS) to inform MEPS respondents' reporting behaviors. MEPS is a nationally representative panel survey studying health care use, access, expenditures, source of payment, insurance coverage, and quality of care. Each year a new panel begins and each panel has 5 rounds of data collection over 2 ½ years that cover a two-year period. Because of a desire to improve round 1 response rates, a decision was made to start the round 1 field period several weeks earlier for the panel beginning in January, 2011; this change lengthened the round 1 field period to 26 weeks (from 23 weeks). While this may have increased the round 1 field period and round 1 response rates in 2011, it also shortened the average round 1 reference period and lengthened the round 2 reference period for some respondents. By starting the interviews of round 1 cases before the interviews of the other rounds, the interviewers could focus exclusively on the more difficult round 1 interviews in the beginning weeks of the round 1 field period. The goal of this paper is to try to tease out whether the change in length of the reference periods for 2011 may have affected MEPS respondents' reporting behaviors. Because MEPS uses the NHIS, conducted by the National Center for Health Statistics, as its sampling frame, both NHIS and MEPS variables will be used in our model to predict MEPS respondent reporting behaviors. Data used will be from the 2010 NHIS matched with the 2011 MEPS files along with additional paradata.

Prepaid Incentives in ABS Surveys: Effect on Nonresponse and Measurement Errors

Meghan McQuiggan[1]*, Rebecca Medway*[1]*, Mengmeng Zhang*[1]*, Mahi Megra*[1]
American Institutes for Research[1]

Extensive literature documents prepaid incentives' ability to increase survey response rates. However, there is less evidence available as to their effect on nonresponse error and measurement error. Do incentives reduce nonresponse error by leading the types of people who tend to be underrepresented in surveys to participate at a higher rate, or do they increase nonresponse error by simply leading more of the same types of people who already tend to participate to take part? Do incentives increase measurement error by leading less interested or motivated individuals to take part, who ultimately provide lower quality data – or do they have little impact on respondents' actions beyond the point of agreeing to participate? Does the answer to these questions differ depending on what kind of incentive is provided? This presentation will utilize the results of a recent incentive experiment that was part of a nationwide ABS field test survey to explore the answers to these questions. In this experiment, sampled households were randomly assigned to one of the following conditions: $5 prepaid + magnet, $5 prepaid only, magnet only, or no incentive. The cash incentive significantly increased the response rate, while the magnet incentive did not. We will explore the effect of the incentives on nonresponse error by comparing the key survey responses and demographic characteristics reported by survey responses in each condition, as well as comparing the distributions of frame variables for respondents in each condition to those of the eligible sample. We will also explore the effect of the incentives on measurement error by comparing the quality of the responses received from the respondents in each group by examining the prevalence of indicators such as item nonresponse, straightlining, and skip errors. This presentation will help researchers considering the use of incentives to be aware of the broader impact that they may have on survey error beyond their effect on the response rate.

How Much is Too Much? Considering the Impact of Survey Response Burden on Nonresponse and Measurement Errors

Mahi Megra[1], Mengmeng Zhang[1], Danielle Battle[1], Rebecca Medway[1]
American Institutes for Research[1]

This presentation will review the results of an experiment embedded in a recent national ABS field test survey that varied the response burden placed on the sampled households by varying the number of questionnaires that sampled households were asked to complete. Once a screener phase established the presence of eligible individuals within the household, some of the eligible households were randomly assigned to receive a single topical questionnaire, while others also received a second topical questionnaire on a different, but related, topic as part of the survey mailing. Sending a second questionnaire to the household did not have a negative impact on the response rate, making it an attractive option to consider for gaining efficiency in future administration by reducing the number of households that need to be sampled in the topical questionnaire phase. Before implementing this strategy in a full-scale administration, it is important to consider whether sending a second topical questionnaire had an impact on who responded or on the quality of their responses. This presentation will explore whether the experiment had an impact on nonresponse bias by comparing key estimates and demographic characteristics reported by respondents from single- and dual-questionnaire households, as well as comparing the distributions of frame variables among single-questionnaire respondents and dual-questionnaire respondents as compared to the eligible sample. It will also explore whether the amount of burden placed on the household had an impact on the extent of measurement error in the provided responses by comparing the prevalence of response quality indicators, such as item nonresponse, straightlining, and skip errors in the single- and dual-questionnaire households.

Errors in Web Surveys as the Main Mode

Session Chairs: Silvia Biffignandi & Fanney Thordottir

Mixed Mode Design with Web Component in Longitudinal Studies

Annamaria Bianchi[1], Silvia Biffignandi[1] , Peter Lynn[2]
University of Bergamo[1] , University of Essex[2]

Mixed mode data collection methods with a web component are increasingly considered as a possibility by many organizations. The inclusion of web into a mixed mode design has potentials to both reduce costs and improve quality. The opportunities for mixed mode data collection with web are particularly appealing for longitudinal surveys, where many information on sample members (e.g. mail address) are known (after recruitment). However, several issues may arise when using web and mixed modes for data collection. One important issue is related to non-response. Response rates are usually very low for web surveys. The inclusion of web into a mixed mode design is expected reduce the risk of error due to non-response. However, the overall effect is not completely clear. Response behavior may be different for different groups using different modes. This may compromise data quality. The aim of this paper is to study the effect of mixing modes with web component on non-response and attrition in longitudinal surveys, both overall and with reference to different groups. The analysis is carried out with reference to the Understanding Society Innovation Panel (IP). The IP is a longitudinal panel designed explicitly to enable methodological research. At Wave 5, a randomised experiment was carried out. One part of the sample was approached face-to-face, while the other part was first invited to complete the survey online and after two weeks people not responding to the online survey were re-approached face-to-face (Burton, 2013). In order to study the impact of mixed modes on attrition, the same design was repeated at Wave 6 (Tarek Al Baghal, 2014). We investigate whether there was a differential attrition effect in the two experimental groups. Interest lies in the identification of subgroups that show differential attrition and non-response effects. Focus is not only related to response rates, but also to non-response bias. The effect of mixing modes is studied with respect to sample composition and survey estimates too. The use of a longitudinal panel allows the availability of a wide range of measures for each sample member, which provide a rather unique opportunity to identify many characteristics of respondents. Also, it allows to investigate the consequences of modifying modes of data collection over time. Findings from this study are relevant both for the development of better methods for longitudinal studies but also for general web-based surveys methodology. References Tarek Al Baghal (2014), Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments, UKHLS working paper 2014-04, ISER, Colchester, University of Essex. Burton, J. (2013), Understanding Society Innovation Panel Wave 5: Results from Methodological Experiments, UKHLS working paper 2014-04, ISER, Colchester, University of Essex.

Minimizing Errors in Multi-mode Surveys Through Adaptive Survey Designs

Joep Burger[1], Koen Perryck[2], Barry Schouten[2]
Statistics Netherlands[1], Statistics Netherlands and Utrecht University[2]

Web surveys have evident benefits over traditional interviewer modes, but also suffer from various sources of error including coverage, nonresponse and measurement effects. These mode effects differ between subpopulations that can be defined by auxiliary information from register data. By applying different strategies to different subpopulations, adaptive survey designs attempt to minimize the mode effects given budgetary and other constraints. Tailoring the survey design is aimed to prevent rather than cure some of the survey errors. We will illustrate how adaptive survey designs can be developed, based on the Dutch Labor Force Survey of 2010 through 2012 and the Dutch Mobility Survey of 2010 through 2013. For both surveys, we have defined a set of potential strategies, stratified the population into relevant groups, estimated input parameters for all combinations of strategy and group, and constructed the optimal design. Through sensitivity analyses we will also show the robustness of the optimal solution against uncertainty about input parameters. A final issue that will be discussed is the robustness of the designs against parameter changes over time, such as declining response rates.

Measurement Error and Nonresponse Error in a Mixed Mode Labor Force Survey

Anton Karlsson[1]
Statistics Iceland[1]

The current situation for many National Statistical Institutes (NSI's) is that while response rates for social surveys seem to be decreasing and the cost efficiency of the data collection phase is being emphasized, users of official statistics still have a need for prompt, accessible, reliable and accurate data. NSI's are therefore looking for ways to increase the number of respondents in social surveys, while keeping the cost of data collection at a minimum in order to be able to provide data fit for use. One proposed solution for this problem is to offer more than one mode in the data collection phase with the possibility that more will respond and, possibly, that non-response bias will be minimized as a result of gaining participation from sample units that would not have responded if a single mode survey would have been fielded. The main problem of this approach is the possibility of measurement error differently affecting responses by the data collection method used. This presentation describes a split ballot test of collecting data for the Icelandic Labor Force Survey using a telephone interview, a web-questionnaire or mix of both modes. The main goal is to examine to which extent different modes can reduce possible non-response bias, without increasing the likelihood of measurement error in the data. Four types of analysis were applied to get a fuller understanding of the effects of offering different modes for the Icelandic Labor Force Survey: 1) The measurement invariance between the three groups was assessed to check if the data collected was comparable across the groups; 2) Differences between measures of key variables from the three split ballot groups were compared use the true LFS as a benchmark; 3) The R-indicator was examined for each of the three groups, as well as the true LFS to to assess to what extent non-response bias might be countered with using different modes; 4) Relative bias was examined for each of the three groups and compared with the true LFS to assess if different modes would succeed in delivering data with less (or more) relative bias than the true LFS. The results suggest that data collected with different modes was mostly comparable and that estimates of key variables were similar across the groups. While the response rate was lowest for the web-only group, its R-indicator was the highest, rivalling the value of the R-indicator of the true LFS.

A Comparison of Errors in Web Surveys Completed Through PC and Mobile Devices

Melanie Revilla[1], Daniele Toninelli[2], Carlos Ochoa[3]
RECSM, Universitat Pompeu Fabra[1], University of Bergamo[2], Netquest[3]

It was observed that some respondents already try to complete web surveys via mobile devices, even when this is unintended. However, we can expect an effect of the device used to complete the survey on the answers and their quality. Indeed, the devices vary at several levels: size of the screen, kind of keyboard, place and conditions (presence of other persons etc) in which the respondents use them, and so on. Therefore, in order to reduce measurement errors and increase comparability of the data collected through different devices in web surveys, it is necessary to study more deeply the differences in errors that appear when respondents answer through PC or mobile devices. In this presentation, we will report the results of an experiment conducted in Spain with the online fieldwork company Netquest, in which the same respondents participated twice to the same survey, once by PC and once by smartphone (optimized or not optimized smartphone versions). In our experiment we also considered two control groups: the respondents of these groups participated to both surveys using the same device.

Data Harmonization

Session Chair: Alan Karr

Using a Total Survey Error Checklist to Investigate Estimate Dissimilarity: Applications with the California Health Interview Survey

Matt Jans[1]
UCLA Center for Health Policy Research[1]

A question often posed to survey methodologists is "Why does Survey A have a different estimate of X than Survey B?" This question captures the motivation for the study and practice of survey methodology and the Total Survey Error (TSE) paradigm as a whole. Survey design and estimation is a “black box” to many researchers, and even trained survey methodologists can find answering such questions to be difficult because numerous design and analysis features can differ across surveys. Thus, a tool can be helpful to avoid overlooking possible error sources, tracking sources reviewed, and weighing the potential impacts of various TSE components. This talk will present a checklist designed around TSE components (Groves, 1989; Groves et al, 2009). TSE decomposes potential error sources into seven sources that contribute to overall error in survey estimates. Representation errors are separated into coverage error (i.e., the difference between an estimate of interest based on elements present on a sampling frame and the true parameter value in the target population), sampling error (i.e., the difference between that estimate calculated on sampled units versus the entire sampling frame), nonresponse error (i.e., the difference in that estimate between units that respond and those that do not respond), and adjustment error (i.e., the difference in that estimate based on adjusted and unadjusted survey data, or the additional error introduced by adjustment). Measurement errors are decomposed into validity (i.e., the difference between the intended construct of interest and the phenomenon actually measured by the question or scale), measurement error (i.e., the difference between what a respondent should report, absent any influence of the interviewer, mode, or respondent confusion or adjustment, and what they actually report), and processing error (i.e., the difference between the value a respondent reports and their value for the same question in a data file). Using two real-world examples from "survey A v. survey B on estimate X" type questions brought before methodological staff of the California Health Interview Survey (CHIS), this talk will discuss the development and use of the TSE Checklist. The checklist has proven useful for organizing the discussion and search strategies in these situations, and is most helpful in situations where resources do not permit new data collection or analysis. While it does not necessarily produce an absolute answer (which are often not possible in such situations), it reduces the number of candidate error sources to a small few that can receive more in-depth investigation. The use of this checklist alone can help educate the people involved about TSE because it emphasizes the complexity of survey data collection and estimation. Application of the checklist can also lead to new insights about survey error and ideas for new analyses or data collection projects in the contexts to which it is applied. These case studies will highlight pros and cons of using the list and its related decision process, and avenues for the future of applied TSE assessment.

The Total Survey Error Framework and the Survey- Quality Controls in the Data Harmonization Process

Marta Kolczynska[1]*, Kazimierz M. Slomczynski*[2]
The Ohio State University, Polish Academy of Sciences[1]*, Polish Academy of Sciences, CONSIRT*[2]

In survey data harmonization, understood as the various procedures by which source variables from existing datasets are combined into a target variables, the analysis of survey data quality is essential both in the process of selecting surveys for harmonization, and for the evaluation of the level of comparability between datasets. We argue that measures of data quality should be included in substantive analyses of harmonized data. In this paper we present an operationalization of the Total Survey Error framework for assessment of survey data quality for secondary data users, distinguishing different stages of the survey process. For our analyses we use several international survey projects, including WVS, EVS, ISSP, ESS, Eurobarometer and its regional renditions, and other well-known studies. Specifically, we apply the following indicators of survey quality: (a) sample drawn from an identifiable frame, (b) information about non-response, (c) indication of any efforts to control the quality of the questionnaire translation, (d) whether there is any indication of questionnaire pretesting, and (e) attempts to control fieldwork. Because of the multi-level structure of data involving these indicators, we propose a statistical model in which the survey is treated as the contextual level of clustering, analogous to models typically used in cross-national analyses. We demonstrate the effects of survey quality controls on the target variables pertaining to trust in public institutions: parliament, political parties and the legal system. The presentation is part of joint projects of the Polish Academy of Sciences and The Ohio State University, supported by grants from the (Polish) National Science Centre (2012/06/M/HS6/00322 and 2012/05/N/HS6/03886).

Discrepancies in Self-Report Diabetes Survey Questions using NHANES, NHIS, and CHIS data

Sarah Lessem[1]
Center for Disease Control and Prevention[1]

Many epidemiological diabetes studies rely on population-based survey data. These studies are based on two assumptions. First, that self-report data is correct and, second, that data quality does not significantly vary across demographic groups. While existing literature indicates systematic error between demographic groups in self-report data, this paper adds to this literature by using internal data inconsistencies to examine the quality of the self-report diabetes data in three health datasets, the National Health and Nutrition Examination Survey 1999-2012 (NHANES), the National Health Interview Survey 1997-2013 (NHIS), and the California Health Interview Survey 2003-2012 (CHIS)(CDC 2014b, a, CHIS 2014). In this article, I illustrate three data discrepancies using what it known about the differences between the epistemologies of type 1 (T1DM) and type 2 diabetes (T2DM) and multiple opportunities to provide the same medication information. I examine three data discrepancies, 1. reporting of having type 1 diabetes and not using insulin, 2. reporting being diagnosed with diabetes at an age which correlates with having type 1 diabetes in almost all cases and not taking insulin, and 3. inconsistent medication reporting. This analysis finds that overall self-reported diabetes data is poor among all demographics with overall rates of data discrepancies ranging from 10.8% to 60.3% depending on the measure. Further, there is a disparity in reporting by demographics with older, less educated, lower income, uninsured, and Hispanic respondents reporting more discrepancies than younger, more educated, wealthier, privately insured and non-Hispanic whites respondents. This indicates that studies relying on these data may misrepresent aspects of diabetes in the population at large and be particularly poor at describing diabetes among underprivileged groups.

Mixed Mode Surveys

Session Chair: Edith de Leeuw

Invited Presentation: **Mixing Modes: Tradeoffs between Coverage, Nonresponse, and Measurement Error**

Roger Tourangeau[1]
Westat[1]

There are at least four reasons to use mixed mode designs in surveys—to improve coverage, increase response rates, lower costs, or reduce measurement error—but there are also a number of potential drawbacks to these designs. This paper will discuss the most common designs for mixed mode surveys, including mix mode designs for cross-sectional surveys (such as the American Community Survey), for longitudinal surveys (such as the Current Population Survey or the National Crime Victimization Survey), and for different components of a single survey (such as the National Survey of Family Growth). In addition, it addresses five questions raised by mixed mode designs: 1) How can we determine the impact of the mode on the responses obtained—that is, the effect of mode on measurement error? 2) How does giving people a choice of modes affect their likelihood of responding? 3) Does the order in which the different modes are offered affect the ultimate response rate obtained? 4) Is it possible to get substantial numbers of cases in a general population sample to respond via the Internet? And 5) What are the advantages and disadvantages of the "unimode" approach to mixed mode design (which attempts to reduce mode effects on measurement) as opposed to the "best practices" approach (which attempts to harness the strengths of each mode to reduce measurement error). Researchers have applied increasing sophisticated models to tease apart the effects of selection on mode differences (that is, the effect of mode on who responds) from the effects of measurement differences (the effect of the mode on the answers respondents give). The paper will review these developments. In addition, it will review the evidence about whether offering people a choice of modes actually lowers response rates. It will also examine the issue of whether beginning data collection with less expensive modes that tend to get lower response rates (e.g., mail) lowers the final response rates relative to starting with more expensive, higher response rate modes (face-to-face). Does having refused to take part in one mode make respondents more likely to refuse in later modes? Next, the paper will examine successes in getting members of the general population to respond to censuses and surveys via the Internet; for example, it will describe the steps Stats Canada took induce a majority of Canadians to complete the 2011 Canadian census on-line. Finally, the paper will try to clarify the tradeoffs between comparability and accuracy in mixed mode surveys. The key considerations are whether the major estimates are objective (rather than attitudinal) and focus on population characteristics (rather than subgroup comparisons).

Invited Presentation: **Mixed Mode Research: Issues in Design and Analysis**

Joop Hox[1]
Utrecht University[1]

Surveys increasingly use mixed mode data collection (e.g., combining face-to-face and web) because this helps to control costs and to maintain good response rates. However, a combination of different survey modes in one study, be it cross-sectional or longitudinal, can lead to different kinds of measurement errors. For example, respondents in a face-to-face survey or a web survey may interpret the same question differently, and might give a different answer, just because of the way the question is presented. This effect of survey mode on the question-answer process is called mode measurement effect. One reason to use a mixed mode design is the potential to reduce coverage errors; switching to a different mode may attract different respondents and therefore improve the coverage. As a result, in practice mode measurement effects and differential selection effects are confounded, and it is difficult to estimate each of these separately. This makes adjustment for mode effects difficult, since adjustment for mode measurement effects is necessary, but adjustment for differential selection is not needed and in fact undesirable. This paper reviews three related issues in mixed mode survey: design, diagnosis, and adjustment. First, since diagnosing and adjusting for mode effects is difficult, the first step in a mixed mode survey should be to design the survey in a way that minimizes mode effects. This involves avoiding differences in questionnaire and implementation details in the different modes, but also actively designing for similarity, called unimode design by Dillman. Second, before adjustment for a mode measurement effect is considered, it is necessary to estimate to what extent apparent differences between modes are the result of differential selection of respondents to different modes. This involves modeling the selection process, followed by estimating the mode measurement effect while controlling the selection effect. Third, in the analysis phase, if there are mode measurement effects, these should be adjusted for. Some proposals for adjustment procedures require auxiliary information, which feeds back into the design phase: not only should mode measurement effects be minimized, but also collecting auxiliary information should be taken into account in the design of a mixed mode survey. This contribution aims to review the methodology for design and analysis of mixed mode surveys in a general way, highlighting design and analysis problems and choices. Statistical analysis models will be discussed; actual estimation procedures will be mentioned but not described in detail.

Evaluating Bias of Sequential Mixed-mode Designs against Benchmark Surveys
Thomas Klausch[1], Barry Schouten[2], Joop Hox[1]
Utrecht University[1], Statistics Netherlands[2]

This study evaluated three types of bias—total, measurement, and selection bias (SB)—in three sequential mixed-mode designs of the Dutch Crime Victimization Survey: telephone, mail, and web, where non-respondents were followed up face-to-face (F2F). In the absence of true scores, all biases were estimated as mode effects against two different types of benchmarks. In the single-mode benchmark (SMB), effects were evaluated against a F2F reference survey. In an alternative analysis, a "hybrid-mode benchmark" (HMB) was used, where effects were evaluated against a mix of the measurements of a web survey and the SB of a F2F survey. A special re-interview design made available additional auxiliary data exploited in estimation for a range of survey variables. Depending on the SMB and HMB perspectives, a telephone, mail, or web design with a F2F follow-up (SMB) or a design involving only mail and/or web but not a F2F follow-up (HMB) is recommended based on the empirical findings.

Errors in Establishment Surveys

Session Chair: Carol House

TSE Sources and Abatement in Establishment Subpopulations
Carl Ramirez[1]
U.S. Government Accountability Office[1]

Business surveys can be subject to unique instances of measurement and particularly representation error – coverage and sampling can be affected by frame quality, multi-unit, multi-level and other corporate structure characteristics. Shortcomings in identifying, contacting and gaining cooperation from respondents who qualify to report on behalf of sampled establishments (or multiple establishments) may outweigh other sources of respondent-driven measurement error. However, working with small samples of specialty business populations may offer the researcher resources not available in larger household or general population surveys. Mixed-mode surveys conducted in 2014 of medical practitioners and drug industry establishments licensed to distribute, administer or prescribe controlled substances in the U.S. showcased specific sources of error and the survey design steps taken to measure and mitigate them. Combining information from auxiliary data associated with list-frame samples, varied advance and followup contacts throughout the survey lifecycle, and interaction with industry-level stakeholders produced insights into and methods to abate business survey error.

Responsiveness and Representativeness in an Establishment Survey of Manufacturers

Eric Fink[1], Joanna Fane Lineback[1]
U.S. Census Bureau[1]

Response rates, because of their ease of calculation and understanding, traditionally have been used as data-collection-quality metrics. However, recent research has cautioned against solely relying on response rates, as survey programs' goals to increase these rates may lead to increasing the likelihood of biasing survey estimates. R-indicators have been proposed as a corresponding measure that can give insight into the data collection process that response rates alone cannot explain (Schoutten and Cobben 2007). In this paper, we calculate traditional response rates and R-indicators for the 2011 Annual Survey of Manufactures and demonstrate that when used in conjunction with each other they can give a more complete picture of the data collection process, particularly the nonresponse follow-up. in particular, we show that despite increasing response rates during the nonresponse follow-up, representativeness across important design variables such as establishment size decreases, owed in part, we hypothesize, to concentrating follow-up on those establishments expected to contribute the most to total estimates. This lack of representativeness is a possible source of bias in resulting survey estimates if nonresponse adjustments do not correct for over of underrepresented areas. Hence, we also examine the R-indicator post-nonresponse imputation. We discuss the tradeoff of reducing sampling variability versus reducing nonresponse bias. Further, we incorporate associated costs into our analysis, and discuss how these cost/quality indicators can be used in conjunction with data quality metrics to provide a more complete picture of the efficacy of the survey process.

A Total Survey Error Approach to Business Surveys
Ger Snijkers[1], Gustav Haraldsen[2], Li-Chung Zhang[3]
Statistics Netherlands[1], Statistics Norway[2], Statistics Norway/University of Southampton[3]

The Total Survey Error approach involves identifying the steps in the survey process, and for each step identifying the resulting survey components and their quality considerations and quality levels. The life cycle process model described by Groves et al (2004) has been used for this purpose. This may work well for social surveys; business surveys, however, are different. In our view business surveys call for a total survey approach built on an expanded process model, which will be discussed in this paper. Snijkers (2015; based on Snijkers, Haraldsen, Jones and Willimack, 2013) has expanded the survey life cycle by including e.g. the planning process with its related error sources. He also pointed out sources of errors linked to the commonly known survey components in business surveys, both at the representation and measurement side of the life cycle. Both the representation and measurement cycle is affected by the fact that business respondents act as informants collecting information and answering questions about defined parts of the businesses which may not coincide with how businesses and business records are organized. In business surveys, unit and measurement issues are interwoven, but we think they can be untangled if we expand the life cycle model into two steps the way Zhang has done in his data collection and transformation model (Zhang, 2011). His model is more relevant to business surveys for two reasons. First business data collections are often designed with a mix of surveys and data capture from administrative registers. Secondly, as Haraldsen (in Snijkers et al., 2013) points out, the transformation step of Zhang's model seems better suited for retrieval activities and judgments taking place in business surveys than the traditional cognitive processes described by Tourangeau in social surveys (Tourangeau et al., 2000). Both these points need to be elaborated. Business surveys are typically self-administrated. Information retrieval often involves an internal data collection and processing before the questionnaire can be completed. This, together with the fact that businesses commonly are sampled for several surveys and that business surveys often are panels, make the administrative tools offered by the surveyor as important for the survey quality as the questionnaire itself. Hence, characteristics of these tools need to be included in a total survey approach. Another important difference between business and social surveys is response burden. In business surveys this is not only a cognitive burden, but also a monetary cost, and may affect quality. Taking a survey error/survey cost approach this means that the time it takes businesses to complete surveys should be included in a cost efficiency model which relates total survey error to total survey costs.

Applying TSE Framework to Comparative Surveys I
Session Chair: Lin Wang

Invited Presentation: **A Total Survey Error Perspective on Comparative Surveys**
Beth-Ellen Pennel[2] , Kristen Cibelli Hibben[2] , Lars Lyberg[1], Peter Mohler[3], Gelaye Worku[1]
Stockholm University[1], University of Michigan[2], University of Mannheim[3]

Surveys that are comparative in nature, i.e., surveys that are multicultural, multiregional or multinational (3M), have an error typology that differs from mono surveys. Even though comparative issues can occur in any survey, 3M surveys typically aim at comparing different populations. Examples of such surveys are the European Social Survey conducted in 20+ countries, the World Values Survey conducted in 75+ countries and the Gallup World Poll conducted in 160 countries. There are a myriad of other 3M surveys in the academic and private sector. Even the European Statistical System can be considered a 3M system since countries conduct studies in accordance with regulations and other instructions issued by the central authority Eurostat. All error sources present in mono surveys are also present in 3M surveys but magnified. There are also operations that are added in a 3M context, such as adaptation of questions so that intended meanings are preserved across, say, nations. Translation per se presents a tremendous problem that might affect respondent comprehension. The most prestigious of 3M surveys have quality assurance and quality control programs based on current best practices. At the other extreme we have 3M surveys that have very little in place that can be characterized as enhancing quality. Some 3M surveys just hand out a source questionnaire and invite interested nations or nation representatives to conduct the survey and it is assumed that the local survey organization takes care of all the details with rather limited instructions. Some 3M surveys have enormous political impact. A recent example is PISA, OECD's international study on student assessments, which ranks countries based on the assessments obtained. The results might very well lead to political reforms despite the fact that the many design and data quality problems should make decision-makers abstain from doing that. There are also examples of data falsification due to financial and national pride issues. The very purpose of comparative surveys is to achieve comparability. An estimate for one country should be reasonably comparable to that of another. But comparability is to a large extent dependent on the various error sources that we are used to, plus some more as mentioned. In this paper we will provide a TSE typology for 3M surveys and give examples of QA/QC approaches that are practiced by some prominent survey organizations and how requirements and standards are developed and implemented. We will also point to resources available in terms of best practice documents, standards and important ongoing research. We will also discuss the pressing problem that most 3M surveys are conducted without any total survey error perspective.

360^0 Quality: Fitness for Use, Total Survey Error, Comparative Error and Survey Process Quality

Brad Edwards[1], Wendy Hicks[1]
Westat[1]

For decades total survey error (TSE) has been the dominant framework for understanding survey quality. Recently the TSE paradigm has been challenged on two fronts: (1) new work on surveys in multinational, multiregional and multicultural contexts (3MC) has adopted a project life cycle approach to process quality assessment and improvement that harkens back in more direct ways to the study objectives; and (2) with rising cost pressures, "fitness for intended use" – the goodness of fit between the study objectives and products – has been offered as a broader framework for assessing survey quality. This paper integrates the TSE paradigm with 3MC and goodness of fit. It offers practical examples of tradeoffs between statistical and comparability errors on the one hand, and factors of cost, relevance, timeliness, and accessibility on the other. Most of the TSE literature is based on cross-sectional, mono-cultural surveys. Recently, some have proposed comparative error in cross-cultural or multinational surveys as another error type (Smith 2011; Edwards and Smith 2013). The literature on comparative survey methods has developed a paradigm for understanding these quality issues (Lyberg and Stuckel 2010; U. of Michigan 2012). The framework borrows from the literature on process quality and project life cycle concepts, initially developed for businesses and manufacturing. The 3MC framework overlaps with the TSE paradigm, but has some key differences. In comparative studies, the focus is on how specific errors affect the analysis of the data as the end product, rather than how they affect the data per se. This enables the comparative quality framework to add another component that relates back to the initial survey design objectives. In contrast, the TSE paradigm has been useful in defining distinct components of total survey error, and in conceptualizing survey error writ large as the difference between the true value in a population and the value as measured by the survey, the mean squared error. But is the whole nothing more than the sum of the parts? Attempts to apply TSE to specific problems in the course of a survey are rare, and TSE's relationship to other quality approaches has not been fully developed. Fitness for use goes a step further, to focus on the user. Biemer and Lyberg (2004) suggest this understanding of quality is a necessary addition to the TSE paradigm. Dimensions of quality from the user perspective may include comparability (as in the 3MC approach), coherence, relevance, accuracy (traditionally the domain of TSE), timeliness, accessibility, and interpretability. Cost, time and available resources mediate all of these dimensions. Unlike TSE, fitness for use can be applied to non-probability surveys (Baker et al. 2013), a fast growing segment of the survey world as probability surveys cope with rising costs and falling response rates. Managing survey process quality and continuous quality improvement have a long history that predates the 3MC work. It is concerned not just about the product, but also the process of making it and the

organization or management of the process. (Lyberg and Biemer 2008) This approach has the potential to unify users and survey practitioners. With quality defined as the degree of fit between a product and its use, it is possible to monitor each step of the survey process to identify non-random errors and address them, achieving an increasingly better product. This paper articulates the relationship of TSE to other paradigms that address survey quality, develops an overarching framework that embraces them in quality profiles, and elaborates the framework with practical examples drawn from recent U.S. national and multinational surveys.

Uncovering Different Sources of Error through Cognitive Interviews in Cross-Cultural Contexts

Alisu Schoua-Glusberg[1]
ResearchSupport Services Inc.[1]

Cognitive interviews have been defined as "the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends." (Beatty & Willis, 2007) While perhaps the most valuable thing we can gather from cognitive testing is the elicitation of question interpretation patterns (do respondents interpret the question as intended, do different demographic and cultural groups interpret the question the same way, etc.), analysis of cognitive interview narrative can help us identify different types of error. When cultural issues get in the way of a group of respondents ability to interpret a question as the designers intended it, a number of mechanisms may play a role in the production of a codable response. The respondent may try to find a response choice that fits their reality, even if the latter does not quite fit in the answers provided. The interviewer may try to 'translate' the respondent's answer into codeable response. Providing examples from cognitive testing of education level questions collected over the last decade, this presentation will focus on different types of respondent error due to interpretation issues with response categories, or to issues in the interaction with the interviewer, and on interviewer error due to incorrect assumptions about the respondent.

Coverage and Nonresponse

Session Chair: Mandi Yi

Invited Presentation: **The Coverage-Nonresponse Trade-off**

Stephanie Eckman[1], *Frauke Kreuter*[1]
RTI International[1], *Joint Program in Survey Methodology, University of Mannheim, IAB*[2]

Undercoverage plaques many types of sampling frames, but fortunately several methods of repairing undercoverage have been developed and tested in the literature. For example, housing units may be missed by listers or may not appear on administrative lists, but we can address this undercoverage with various forms of the missed housing unit procedure (Kish 1965, Section 2.8). Persons with tenuous connections to households are sometimes not captured in rosters, but detailed probes about household members can improve coverage (Martin 1999). Respondents may try to hide their eligibility during screener interviews, but screener questions that disguise the target population can increase eligibility rates (Tourangeau et al 2013). However, there is evidence from diverse sources that the cases identified and added via efforts to improve coverage are disproportionately nonresponders to the survey request. For example, the AAPOR Cell Phone Task Force found that mobile-only households, which are undercovered in landline frames, have lower response rates than households which have a landline (AAPOR Cell Phone Task Force Report). The LISS web panel offered internet access to non-internet households, in an effort to improve coverage of the population, but observed lower recruitment rates among these households (Leenheer & Scherpenzeel 2013). Because response rates are published and seen as quality indicators, whereas coverage rates usually go unreported, there may be incentives for those involved in survey production to increase response rates at the expense of coverage rates. This chapter will systematically review the existing evidence for such a nonresponse - coverage trade-off and use a theoretical lens to search for the mechanisms underlying the connection between nonresponse and undercoverage. We will also call attention to situations in which the distinction between nonresponse and undercoverage is not entirely clear. For example, the literature on web surveys tends to call persons without internet access undercovered, even though they could be, in some designs, counted in the denominator of the response rate. We challenge the TSE framework's distinction between these error sources, and consider alternative formulations of the response rate that collapse across response and coverage. Such alternative formulations will be particularly important as the field moves towards data collection beyond surveys as we know them, where nonresponse and undercoverage cannot be easily distinguished.

Estimating Percentages or Proportions in the Presents of Undercoverage and Nonresponse

Robert Tortora[1]
ICF International[1]

Coverage bias or error can occur when 1) part of the target population is not accessible through the sampling frame, 2) when the frame contains (undetected duplicate units and 3) when the frame contains out of scope units. In this paper I examine the potential coverage bias when part of the population is not accessible through the sampling frame. This would happen, for example, if newly registered voters were not included in a sampling frame of registered voters. Estimating percentages or proportions provides an opportunity to examine the impact of undercoverage since these estimates are bounded between 0 and 100 (0 and 1) for percentages (proportions) in contrast say to estimating household income which has a larger an unknown range. If PC represents the estimate from the covered portion of the population and PNC represents the (unknown) estimate that would come from those units not in the sampling frame the estimate of the overall percentage is given by P = (NC/N)PC + (1 - NC/N)PNC where NC/N is the proportion of the population covered by the sampling frame. For this paper assume a simple random sampling of size n. We are concerned with identifying the values of NC/N and PNC such that P still belongs to the (1 – α) % confidence interval for PC, that is where we do not have to be concerned with introducing bias into the estimate. Since PC, PNC and NC/N are bounded varying these quantities in a simulation can identify when P is contained in the confidence interval for PC for various sample sizes and values of α. Smaller sample sizes and lower values of α improve coverage for P. In addition the risk of P being outside of the confidence interval is examined by assuming PNC has is drawn for a probability distribution. We examine two cases, when PNC is drawn for a uniform distribution and a normal distribution. These results are extended to incorporate nonresponse since PC can be expressed as a linear combination of the estimate for respondents and the unobserved estimate for nonrespondents.

Total Survey Error: Sample Error Combined with Coverage and Non-response Error

David Rothschild[1]
Microsoft Research[1]

In this paper we examine the interplay between: coverage, non-response, sample, survey, time, and herding bias and/or error. Our dataset is all publicly available "probability" surveys that are listed on major survey aggregation sites between 1998 and 2014 for presidential, senatorial, and gubernatorial elections. We determine the overall error by election, election-type, and cycle. We demonstrate that there is a consistent coverage and non-response error, on top of sample error, that raises total survey error above the sample error reported by major survey firms. We are able to separately identify the impact and scale of survey error. Finally, we do not find conclusive evidence of herding.

A Validation Study on Voter Turnout Bias in Switzerland

Ben Jann[1]*, Simon Hugi*[1]
University of Bern[1]

"Who's voting?" is an important research question in political science as coverage error because politically inactive people are more likely to be excluded from the sampling frame. Second, non-voters might be more likely to refuse participation in election studies, leading to nonresponse bias. Third, there might be measurement bias because respondents over report their political participation (e.g. due to social desirability of voting).
To disentangle the three sources of bias we conducted a validation study. The study was carried out in a small town in the region of Bern, Switzerland, in the aftermath of the popular federal vote of September 2013. We first sampled 2000 citizens from the electoral register of the town and then collected directory-listed landline numbers for the households of the sampled citizens to mimic a typical CATI sample. We then contacted the 1679 citizens whose households were listed in the telephone directory for an interview, of which 893 participated in the survey. After conducting the interviews we matched the records of the 2000 citizens to the voting cards archived at the town's administration and linked the records to some further administrative data sources.
In line with international research, our results show that all three sources of error inflate the survey estimate of the turnout rate – in our case by a total of 20 percentage points. However, coverage error, primarily caused by under coverage of young people with low SES, is only moderate compared to the more pronounced biases due to nonresponse and over reporting. For example, over 25% of validated nonvoters falsely claimed in the survey that they had voted. Our findings also demonstrate that the biases are difficult to approach by weighting schemes and that they do affect the results from participation models. In our survey we also included a wording experiment to reduce over reporting (Belli et al. 1999). The revised wording, however, did not yield more valid results than the standard wording.

3:30 - 5:00 p.m.
Paper Session IV

Nonresponse and Measurement Error II
Session Chair: Nancy Clusen

Adjusting for Measurement Error and Nonresponse in Physical Activity Surveys: A Simulation Study
Nicholas Beyler[1], Amy Beyler[2]
Mathematica Policy Research[1], UnitedHealthcare[2]

Adult Americans are encouraged to engage in at least 150 minutes of moderate to vigorous physical activity (MVPA) each week to improve and maintain their health. National surveys which collect physical activity data to assess whether or not adults adhere to this guideline use self-report questionnaires which are prone to measurement error and nonresponse. Studies have examined the individual effects of each of these error sources on estimators of physical activity, but little is known about the consequences of not adjusting for both error sources. Using a model-based approach we conduct a simulation study to determine how estimators of adherence to the guideline for adults to engage in 150 minutes of MVPA each week respond to different magnitudes of measurement and nonresponse errors in self-reported physical activity data. We consider the biases in estimators which account and adjust for measurement and nonresponse errors, measurement errors only, nonresponse errors only, and neither error source. Estimators that adjust for both measurement and nonresponse errors provide the least amount of bias across all simulation scenarios. In some scenarios the naïve estimator, which does not adjust for either error source, results in less bias than estimators that adjust for only one error source. To avoid biases when estimating physical activity outcomes from national surveys, adjustments for both measurement error and nonresponse should be considered. Survey designs for studies that include physical activity self-report questionnaires should allow for such adjustments.

Invited Presentation: **The Effect of Nonresponse and Measurement Error on Wage Regression Across Survey Modes: A Validation Study**
Antje Kirchner[1], Barbara Felderer[1]
University of Nebraska[1] , University of Mannheim[2]

In order to draw valid conclusions from survey interviews (e.g. population means or regression parameters) it is important that the data is of good quality, e.g. not biased by nonresponse or measurement error or affected by the collection mode. To compare nonresponse and measurement error bias across telephone and web modes, we use administrative records and survey data. In an experimental setting we randomly assigned respondents to either telephone or web mode (n=3,482). Because the sampled persons were selected from German administrative records, record data are available for all sample units to study nonresponse and measurement error at the same time. We find differential nonresponse bias and measurement bias for univariate mean statistics for substantive survey variables such as income and receipt of unemployment benefit during the last 12 months for both survey modes, and less in socio demographic characteristics. However, the question we focus on in this paper is how these errors affect multivariate analyses at the micro level, e.g. regression coefficients. In each mode, we run traditional wage regressions--regressing monthly income from employment on the type of employment, gender, age and an indicator of whether the respondent was unemployed during the past 12 months. Earlier analyses show that several of these variables are subject to nonresponse and measurement error. For each mode, we thus run the model with three sources of data: administrative data for the gross sample (nonrespondents and respondents) administrative data for respondents only survey data for respondents The benchmark estimate to evaluate measurement error and nonresponse is given by Model 1. The comparison between Model 1 and 2 allows us to evaluate the effect of nonresponse bias on the regression parameters, whereas the comparison of Models 2 and 3 reveals the effect due to measurement bias. Finally, comparing Models 1 and 3 reveals the combined effect of nonresponse and measurement error on the regression coefficients. Finally, Model 3 is rerun for both modes using different nonresponse adjustments based on socio-demographics available on the frame. Adjusting for nonresponse should decrease nonresponse bias but it is less clear which effect this will have on bias due to measurement error and thus on the overall bias. Weighting could increase bias if there is a correlation of response propensity and response quality e.g. people with low propensity give low quality answers. These cases would be given high weight in the nonresponse adjustment, which could seriously affect the estimated regression coefficients. This paper will discuss how regression coefficients are affected by the biases we found examining univariate survey characteristics. We will demonstrate which error source exerts the larger effect and whether the errors reinforce each other or cancel each other out. Finally, the paper

assesses whether the effects are the same for both modes, or whether there is a 'preferred' mode. The results in this paper will help to understand the interaction of nonresponse and measurement bias and the consequences of nonresponse adjustment on the combined error.

Is More Always Better? The Impact of Adaptive Design on Nonresponse and Measurement Errors

Jaki McCarthy[1], Tyler Wilson[2], Andrew Dau[2], Kathy Ott[2]
US Department of Agriculture/National Agricultural Statistics Service[1], USDA/NASS[2]

NASS has recently begun using adaptive design approaches to strategically plan data collection in the Agricultural Resource Management Survey (ARMS). The operational data collection strategy for ARMS includes multiple mail outs of the questionnaire, followed by a field interviewer follow up for mail nonrespondents. We have developed nonresponse propensity models for ARMS that can be used to identify records likely to be nonrespondents. As part of testing adaptive design approaches we identified records with the highest likelihood of nonresponse (70% or higher). Special contact procedures were developed for these difficult records (Mitchell, Ott and Ridolfo, 2014). A second set of records with nonresponse propensities between 50 and 69% was also targeted. This set of records is the subject of this paper. Our adaptive design excluded these records from the questionnaire mailings. Instead, field enumerators were provided with a questionnaire package including the questionnaire, cover letter and information about the survey, token items bearing the NASS logo, and an interviewer contact card. Interviewers were instructed to contact the sampled operations in person to gain cooperation. They then left the questionnaire package with the respondent to complete and scheduled a time to return to collect the completed questionnaire. Response rates for this group will be compared to a control group receiving the traditional mail with field interviewer follow up procedures. While the intent of these procedures is to increase unit response rates they may also impact data quality. We will examine item nonresponse rates, item edit rates and other quality indicators and compare them across the two groups. It is possible that while response rates increase by using a drop off/pick up procedure thus allowing respondents to complete questionnaires with little to no assistance from interviewers, the quality of the data reported may change. Does this procedure allow more errors? Are skipping patterns or item nonresponse affected? Or, does the additional flexibility of when and where respondents complete the survey allow respondents to report more completely and accurately?

A Common Metric for Nonresponse and Measurement Error

Peter Lugtig[1]
Utrecht University[1]

Nonresponse and measurement error are perhaps the two survey errors that receive most attention from survey methodologists. Strangely, we often do not know whether nonresponse or measurement error contributes most to total survey error in any particular survey. When we can assess nonresponse error and/or bias, we use sampling frame or population-level data to do so. Usually, this information is about socio-demographic characteristics of respondents and nonrespondents. Often, these are not the variables we are interested in when assessing measurement error, and as such, we cannot assess whether nonresponse or measurement error is worse for our variable of interest. So currently, the only way to assess the trade-off between nonresponse and measurement error is by having validation data for both respondents and nonrespondents in a survey. In this paper, I propose a method with which the size of nonresponse and measurement error can be estimated and compared for attitudinal variables. The methods works only when: 1) data are collected longitudinally, and one is interested in nonresponse bias that occurs because of attrition (so initial nonresponse bias cannot be assessed) 2) a Multi Trait Multi Method (MTMM) model is used to assess the reliability and validity of the variable of interest in one of the first waves of the longitudinal study. The presentation will show how the traditional MTMM model can be extended to include nonresponse error. Nonresponse error can be expressed in terms of the parameters of the MTMM model, meaning that the size of measurement and nonresponse error can be assessed for 1) means 2) variances and 3) covariances of the variable of interest over the course of the panel study. In the paper I will illustrate the general approach, and in the presentation I will focus on an example using data from the LISS panel that will illustrate the sizes and interactions of nonresponse and measurement error for social trust.

Error Sources in Web Surveys

Session Chair: Mark Schulman

Reducing Measurement Error in Online Surveys Through Usability Evaluation

Lin Wang[1]
U.S. Census Bureau[1]

Online surveys are now widely used to gather information from the public. In addition to various sources of measurement error commonly found in self-administrated survey instruments, such as variation in question interpretation, online survey instruments are uniquely subject to website usability problems. Usability, in this context, refers to the extent to which a respondent can self-administer an online survey effectively, efficiently, and satisfactorily. Examples of usability problems include navigation difficulties, keystroke issues, and legibility issues. Such problems may compromise a respondent's ability to provide accurate responses, frustrate the respondent, and slow down survey administration. Online surveys, as one type of web application, require the same key elements as other types of web application, e.g., organized content, clear labeling, easy navigation, and good accessibility. To minimize measurement-error vulnerability, online survey instruments must begin with and adhere to the principles of human-centered design, which emphasizes understanding the user, tasks and context; involving the user in the design; and addressing the whole user experience. In this study, we will explore how usability evaluation can effectively identify problems that may cause measurement errors. We begin by analyzing respondents' sensory, perceptual, and cognitive capacity for interacting with online survey instruments. Next, we discuss possible measurement errors attributable to the incompatibility between participants' information processing capabilities and the user interface design of online survey instruments. Finally, we introduce a systematic approach to usability evaluation of online survey instruments, that includes evaluation design, data collection, data analysis, and results dissemination.

Web Surveys: Errors in the Process and TSE for a Quality Perspective

Silvia Biffignandi[1], Fanney Thorsdottir[2]
University of Bergamo[1], University of Iceland[2]

Traditionally, literature on survey methodology relies on the Total survey error (TSE) approach for classifying sources of errors in web surveys. The TSE framework is an effective approach for understanding error sources in a comprehensive way. However, a more useful approach to data quality control is to integrate the TSE framework and a process quality perspective. The process oriented approach decomposes the survey into steps that define the flow of the survey process. When the steps have been defined, risks of errors that may arise in the survey process can be taken into account at specific steps in the process. Even though the effectiveness of the process quality approach has been emphasized for surveys in general, integration of the TSE and the process frameworks is rarely adopted in literature on Web survey methodology. In many respects, Web surveys involve different steps and imply different participation behaviours from other survey modes. Thus, errors and risks of errors in Web surveys are for some aspects of the survey process different from traditional surveys. In this presentation the steps involved in the Web survey process are defined and presented in a flow diagram. The diagram enables researchers who carry out research within the field of survey methodology to have a common general framework of the Web survey process and relate their studies of errors to specific steps in that process. This will result in a more effective communication between survey methodologists when they discuss their research findings. The flow diagram is also useful for the survey practitioner who needs to have a clear understanding of the steps involved when a survey is conducted. Thus, the diagram can be used as a practical scheme for organising a Web survey. The paper could fit to a session like : The role of the TSE paradigm in survey management and quality control and quality assurance.

Nonresponse and Measurement Bias in Web Surveys

Anke Metzler[1], Marek Fuchs[1]
Darmstadt University of Technology[1]

Web surveys suffer from substantial nonresponse inducing nonresponse bias in estimates. Various groups cause nonresponse: initial nonrespondents and survey break-offs. The discussion whether or not nonrespondents should be incorporated into Web surveys is commonly known. On the one hand Web survey researchers have established various strategies to reduce nonresponse rates. However, on the other hand it is questionable, whether nonresponse reduction strategies will improve data quality. Respondents, who are less motivated, might take cognitive shortcuts more often while answering questions and provide less accurate responses, thus incorporating these respondents into the net sample might increase measurement error. Therefore, it is important to consider the impact of the nonresponse and measurement error to decide whether or not strategies that help increase response rates should be applied. In the analysis reported in this paper, we used data from three large scale surveys among university applicants conducted in 2012, 2013, and 2014 with approximately 6,000 respondents each. Each survey achieved response rates in the range of 30 to 40 percent. However, regardless of the considerably high response rates analyses revealed significant differences between the gross sample and respondents in the net sample with respect to socio-demographic and other background variables resulting in knowledgeable nonresponse bias. Interestingly, respondents who broke-off differ from initial nonrespondents. Also, more

in-depth analyses of the nonresponse bias indicate that the initial nonresponse bias and the survey break-off bias add to each other but that nonresponse bias is caused by initial nonresponse to a greater extent. Therefore, it seems to be more substantial to incorporate initial nonrespondents. However, it is easier and more cost-effective to convince survey break-offs to participate in a survey by optimizing the questionnaire design. Additionally, results indicate that the substantive answers of the gross sample do not differ from the n et sample. Nevertheless, the measurement error increases, if survey break-offs are incorporated. First results imply that survey break-offs are prone to speeding, read instructions less careful and differentiate less between responses to grid questions in contrast to respondents who completed the questionnaire. Convincing survey break-offs to participate in a survey would reduce nonresponse bias but cause additional measurement bias. Results suggest that survey break-off should not only be considered in terms of the nonresponse bias prevention but also in terms of the measurement error.

Classification and Editing

Session Chairs: Lars Lyberg

Invited Presentation: **Quantifying Measurement Errors in Partially Edited Data**

Thomas Laitila[1], Karin Lindgren[1], Anders Norberg[1], Martin Odencrants[1]
Statistics Sweden[1]

One of the most damaging sources of the nonsampling error component of the total survey error is the measurement error. Measurement errors occur when the observed value differs from the true value according to the definition of the variable (Biemer and Lyberg, 2003). Causes of measurement errors can be the data collection instrument; the mode and the respondent etc. Editing is a major activity to resolve and treat some of the measurement errors, where collected data are reviewed for detection of inconsistencies and errors. The aims of an editing process should be to give detailed information about the quality; to provide for future improvement of the survey; and to tidy up the data (Granquist, 1984). Editing is costly and time consuming. Developments of new theories and methods useful for reducing the resources spent on editing of survey data are therefore of interest. Such developments will also be necessary for treating registers and data sets of sizes implied by the Big Data concept. One approach towards a more efficient editing is selective editing, where only a subset of suspicious data is treated (Granquist and Kovar, 1997). Here the leading idea is to spend resources only on those observations which have potential effects on the estimates. Selective editing is based on the calculations of "global scores", expressing a combined measure of importance in estimation and suspicion of measurement error. Based on these global scores, observations are selected for editing. Selective editing has been developed from practical experiences and lacks a theoretical framework (de Waal, 2014). In particular the effect of selective editing is established on analyses of previously collected data sets, while there is no direct measure of error in the estimates due to measurement errors among unedited observations. One approach to correct for remaining measurement errors is suggested by Ilves and Laitila (2009), where observations selected for treatment are drawn using a sampling design. Global scores are preferably used for sampling, to give more important and suspicious observations a lager probability of selection. A different approach for measuring remaining measurement errors after selective editing is considered in this paper. Editing a subset of observations yields information on the distribution of measurement errors and its dependence on the score constructs used for selection. One way of making use of this information is to adapt a model based approach and regressing observed measurement errors on calculated scores. Since the scores are available also for the unedited set, the estimated model can predict measurement errors in unedited observations. Summing these predictions yields a measure of the effect of remaining measurement errors after selective editing, and gives valuable information on survey quality. The adaption of a model based approach implicitly assumes the measurement errors as outcomes of random trials. This paper suggests a theory making such an assumption reasonable in survey data. The theory also provides with insights on the distribution of measurement errors suggesting relevant statistical models. Data from different surveys at Statistics Sweden are used for illustration of the theory and methodology.

Invited Presentation: **Classification Error in Crime Victimization Surveys: A Markov Latent Class Analysis**

Marcus Berzofsky[1], Paul Biemer[1]
RTI International[1]

Many countries rely on crime victimization surveys to assess the volume of crime and to monitor trends in victimization rates among their populations. Obtaining accurate survey data on crime victimizations is challenging due to variations in how specific types of victimization are defined and understood by respondents, memory errors, sensitivities in reporting some types of crime, respondent burden issues and so forth. Although there have been efforts over the years to improve

the quality of victimization data in surveys, concerns persist that victimizations may be substantially under-reported for some crimes and that crime statistics, generally, are inaccurate. Evidence to support these concerns is usually provided from studies comparing survey reports with official police reports. However, it is well-known that such comparisons do not reflect the true error in the survey reports because, for example, they are confined to victimizations that have been reported to law enforcement and, therefore, may also be more accurately reported in surveys. This paper quantifies the classification error in crime victimization surveys using an innovative application of Markov latent class analysis (MLCA). MLCA will be applied to the largest victimization survey in the U.S. – the U.S. National Crime Victimization Survey (NCVS) – to achieve two objectives. First, we evaluate the classification error in questions that elicit whether or not a certain type of victimization has occurred during the survey reference period – i.e., the so-called screener questions. The error rates for the screener questions will be assessed for specific population subgroups as well as the population as a whole. Like other crime victimization surveys, the NCVS presents a number of important challenges for applying MLCA due to the complex sample design, low prevalence outcomes, and high dimensional data. Thus, the second objective of the paper is to show how these challenges can be overcome in the modeling process using an innovative model fitting approach developed by the authors specifically for MLCA applications. In this paper, we will estimate the classification error rates for key outcome variables in the NCVS, investigate some of the causes and correlates of the errors, and assess the impact of these errors on the publish estimates. The primary contributions of the paper to the study of total survey error are two-fold. First, the paper is unique in that it provides model-based estimates of classification error in the NCVS. Information on measurement errors in crime victimization survey is quite scant in the extant literature and this paper is one of only several attempts at actually quantifying classification error and bias in crime victimization rates. Second, the combination of the complex survey design of the NCVS, high levels of missing data and the low prevalence of crime victimizations creates important difficulties in applications of MLCA to victimization survey data. These difficulties are exposed in the analysis and innovative solutions to overcome these difficulties are provided.

Sampling Design Alternatives for Quality Checks in the Diabetes Prevention Program Outcome Study

Michael Larsen[1]
The George Washington University[1]

The Diabetes Prevention Program Outcome Study (DPPOS) the continued follow-up of the cohort from a multi-center clinical trial of the primary prevention of Type 2 diabetes. The original study, DPP, demonstrated the effectiveness of lifestyle intervention and metformin use in preventing onset of diabetes in an at risk population. DPPOS studies the effects of the interventions on further development of diabetes and diabetes complications, including retinopathy, cancer, and cardiovascular disease. Data are gathered at clinical visits, but also through participants surveys. Strict protocols are used for data entry and preservation of materials. Translation of medical records to paper forms to the study computer database is checked for quality on an ongoing basis. Periodically in-depth checks for missing forms, deviations from good clinical practice, and field errors have been done. This talk presents a sample survey approach to this problem that reduces work load, maintains adequate confidence in quality, and can be better adapted to future paperless data entry.

Effect of Missing Data on Classification Error: An application of Two Full Information Maximum Likelihood Techniques with a Markov Latent Class Analysis

Susan Edwards[1], *Marcus Berzofsky*[1], *Paul Biemer*[1]
RTI International[1]

Sensitive outcomes on surveys are plagued by item nonresponse and measurement error – referred to as classification error for categorical outcomes. Both of these types of error can lead to biased estimates and, potentially, erroneous conclusions if not properly understood and addressed. The National Crime Victimization Survey (NCVS), administered by the U.S. Census Bureau on behalf of the Bureau of Justice Statistics, is a nationally representative rotating panel survey with seven waves of the non-institutionalized United States population which measures two types of crime victimization – property and violent. Because not all crime is reported to the police, there is no gold standard (i.e., error free) measure of whether a respondent was actually victimized. For panel or longitudinal data, Markov latent class analysis (MLCA) is a model-based approach which uses response patterns across interview waves to estimate the false positive (i.e., a respondent reports being a victim when they truly were not) and false negative (i.e., a respondent reports not being a victim when they truly were) classification probabilities. This paper builds on research presented by Berzofsky, Edwards, and Biemer (2014) that showed using missing at random (MAR) adjustments for missing data impacted the latent class analysis (LCA) estimates measuring sexual victimization. Typically, LC and MLC models are fit using complete data; that is list-wise deletion is preformed to remove any observations with item non-response. Recently software, such as LatentGOLD and LEM, has been developed to utilize all records during model estimation using the relatively new

technique of full information maximum likelihood (FIML). In 1982, Fuchs' proposed a FIML method to handle non-response which is MAR; Fay expanded the FIML method in 1986 to handle non-response not missing at random (NMAR) by modeling the response pattern. Due to the rare nature of violent crime victimization, our analysis combines multiple years of NCVS data to achieve adequate precision in the results. Using LatentGOLD, which can account for complex survey designs, we assess the impact that the MAR and NMAR FIML methods for including cases with missing data have on the classification error estimates for violent crime victimization.

Survey Innovations in a Total Survey Error Context

Session Chair: Peter Miller

Examining Mode Options for the Commodity Flow Survey

Joanna Lineback[1], Robert Ashmead[1], Eric Slud[1]
U.S. Census Bureau[1]

In this paper, we examine the cost-quality tradeoffs of the Commodity Flow Survey (CFS) moving from a paper-only data collection to a paper and electronic collection to possibly an electronic-only collection. The CFS is a mandatory survey sponsored by the U.S. Bureau of Transportation and conducted by the U.S. Census Bureau. It is conducted every five years, on a quarterly basis, to provide estimates on the movement of goods in the United States. Historically, the CFS was a single-mode (paper) survey collected by mail. Recently, there have been changes to the data collection methods. The 2012 CFS was a multi-mode survey; respondents were given the option of reporting by paper or electronically, and an all-electronic collection is being considered for the 2017 CFS. To inform this decision, we used response data and a rich set of available auxiliary information, including historical data, paradata, cost data, and administrative records data, to develop cost and quality metrics that were examined in conjunction at key stages of data collection and processing.

Adaptive Design for the National Teacher and Principal Survey (NTPS)

David Marker[1], Lou Rizzo[1], Minsun Riddles[1], Erin Wiley[1], Andrew Zukerberg[2]
Westat[1], U.S. National Center for Education Statistics[2]

Statistical agencies are frequently confronted with the trade-offs between timeliness (relevance) and accuracy. Waiting for the last responses and quality reviews can improve accuracy but delay production of the data sets and analyses, reducing their relevance to users. The National Center for Education Statistics has conducted the quadrennial Schools and Staffing Survey (SASS) since the 1980s. Beginning with the 2015-16 school year SASS will be replaced with a new biennial National Teacher Principal Survey (NTPS). As part of the design for the NTPS, we reviewed response patterns and paradata collected during the 2011-12 SASS to develop an adaptive design for the new study. Adaptations may include when to switch data collection modes, when to stop overall data collection, and revisions to methods for contacting respondents. We are therefore simultaneously examining multiple components of Total Survey Error, including nonresponse bias, mode effects, and relevance (time lag from reference period to publication). This presentation will discuss the findings from the SASS review, how it is being implemented in the 2015-16 NTPS, and provide a framework for other studies considering adaptive design approaches.

Modeling the Effects of Innovation in the National Crime Victimization Survey

Joseph Schafer[1]
United States Census Bureau[1]

Field staff for the National Crime Victimization Survey (NCVS) experienced two major innovations during 2011 and 2012 that may have impacted data quality. The first, a program of refresher training and performance monitoring, was phased in by a randomized experiment. The second, a nationwide field realignment program that reduced the number of Census Bureau regional offices from twelve to six, was phased in without randomization. To estimate the effects of these innovations, we developed two sets of hierarchical Bayesian longitudinal models describing data-quality metrics and survey outcomes at the interviewer level. One set of models relied heavily on the experimental design and observed differences between cohorts; the other set attempted to extract intervention effects by removing the influences of confounding covariates, long-term historical trends and annual periodic cycles. Taken together, they serve as useful case studies for assessing the impact of changes to survey environments when randomization is practical and when it is not.

Error and Cost Tradeoffs Involved in Innovations for Decennial Census Data Collection

Peter Miller[1], Mary Mulry[1], Gina Walejko[1]

United States Census Bureau[1]

Faced with the obligation to contain costs, the United States Census Bureau is investigating the use of multiple modes of data collection for the 2020 Census. The basic protocol followed for decades - a self-response phase followed by face-to-face interviews with non-respondents - will be employed again. Self-response mode options will expand to include a Web questionnaire in addition to the traditional mailed paper instrument. In addition, administrative records may be used to determine whether households are occupied or vacant and to enumerate non-responding households. Web data collection and use of administrative records are aimed at reducing costs entailed in self-response and in nonresponse follow-up. These multiple data collection approaches pose data quality and cost tradeoffs that can be examined through the total survey error perspective. For example, in the self-response phase of census data collection, the addition of a Web response option has the potential to reduce costs and time for enumeration, but coverage and non-response error are likely to be higher for this mode than for mailed paper questionnaires. There can also be measurement error tradeoffs between Web and paper versions of the self-response questionnaire. An instrument that takes full advantage of capabilities offered by Web data collection may produce data that differ substantially from those gathered through paper questionnaires. In the non-response follow-up phase of data collection, there are error and cost tradeoffs entailed in collecting data by interviews with households, by interviews with proxy respondents and by using information from administrative records. Employing administrative records for enumeration can reduce costs markedly, but records do not cover all households and, depending on their vintage and original purpose of the records, the information they contain may be inaccurate or incomplete. Face-to-face interviews, by contrast, may have better coverage and measurement properties, but they are much more expensive. Accepting proxy reports for households can save money but can also introduce more measurement error. In this paper, we discuss these and other cost and quality tradeoffs and review ongoing research to address these issues in preparation for the 2020 Census.

Estimating and Adjusting Survey Errors in Mixed-Mode Data, Part I

Session Chair: Thomas Klausch

Estimating Unemployment Levels Using a Mix of Three Interview Modes

Bart Buelens[1]*, Jan van den Brakel*[2]
Statistics Netherlands[1]*, Statistics Netherlands & Maastricht University*[2]

Combining web interviewing with telephone and face-to-face interviewing is attractive because of its lower administration cost and its potential to reduce selection bias in sample surveys. National Statistical Institutes are introducing mixed mode surveys into their regular programs. The Labor Force Survey (LFS) conducted by Statistics Netherlands is an example. The survey is conducted by sequential use of web, telephone and face-to-face interviewing since 2012. In sequential mixed mode designs the distribution of the respondents over the different interview modes is generally not constant in consecutive editions of a repeated survey. This may cause effects associated with the modes, such as measurement bias, to vary over time. Time series based on repeatedly conducted mixed mode surveys will therefore reflect a more severely biased estimate of changes over time of the variables of interest compared to single mode surveys. Two estimation methods that are robust to variations in the distribution of respondents over the different modes are applied to the monthly LFS. The first approach is based on the general regression (GREG) estimator. Measurement bias between subsequent editions of the LFS is stabilized by calibrating the survey response to fixed distributions over the interview modes (Buelens and Van den Brakel, 2014). The use of this predominantly design based approach is motivated with a measurement error model for the observations obtained in the sample. The second approach uses a linear model to estimate measurement errors and predict individual responses under alternative modes. These predictions are used in the GREG estimator to obtain parameter estimates under the different modes (Suzer-Gurtekin et al., 2012). The two methods are used to produce monthly estimates of the number of unemployed people in the Netherlands. The results are compared; advantages and disadvantages are discussed. Buelens, B. and J. A. van den Brakel (2014). Measurement error calibration in mixed-mode sample surveys. Sociological Methods & Research, first published on May 12, doi: 10.1177/0049124114532444. Suzer-Gurtekin, Z., S. Heeringa, and R. Vaillant (2012). Investigating the bias of alternative statistical inference methods in sequential mixed-mode surveys. In proceedings of the JSM, section on survey research methods, pp. 4711-4725.

Mixed-Mode Inference: An Imputation Approach Incorporating Covariances between Modes

Zeynep Suzer-Gurtekin[1]*, Richard Valliant*[1]*, Steven Heeringa*[1]
University of Michigan[1]

This study extends the imputation approach to incorporate covariances across modes. The current imputation method assumes independence across the models and fits the imputation models separately. The study uses the American Community Survey (ACS) data to evaluate the method empirically.

Estimating and Adjusting Bias of Sequential Mixed-mode Surveys Using Re-interview Data
Thomas Klausch[1], Barry Schouten[2]
Utrecht University[1], Statistics Netherlands[2]

Mixed-mode (MM) designs have become an important method in surveys carried out by national statistical institutes (NSIs) and in the creation of official statistics for large populations. One of the key problems that can occur in MM surveys are so-called measurement effects. A measurement effect is the increase in systematic measurement error of a MM design caused by some mode(s) measuring a target variable with larger error than others. This threatens measurement validity and may offset reductions in selection bias gained by combining modes in the design. In MM inference it is desirable to adjust measurement effects towards a mode which is considered most valid, the so-called measurement benchmark mode. Our statistical adjustment approach regards the MM data as a missing data problem, in which each respondent has either an observation or a potential outcome under the benchmark mode. The benchmark outcome is potential, if respondents are observed under a different (i.e. less accurate) mode used in the design. Subsequently, a missing data technique, such as imputation, is applied to solve the missing data problem by imputing potential outcomes on the benchmark. In this endeavour auxiliary data is needed which allows a so-called 'missing at random assumption'. This assumption suggests that the relative selectivity between modes on the target variable is fully explained by the auxiliary data. In practice, the availability of strong adjustment data is scarce, however. We suggest an approach which aspires to increase the availability of auxiliary data used in missing data adjustment of the potential outcomes by a re-interview of a random sample of respondents in the MM design. The re-interview data is collected regardless of the mode in which the respondent participated. For example, in a sequential MM survey that follows up web nonrespondents in face-to-face, some additional data is collected from the respondents in web by a re-interview in face-to-face. For these respondents, two measurements become available, one in web, and one in face-to-face. This data is used to estimate a measurement model of the true relation of web and face-to-face measurements that is applied in the prediction of benchmark outcomes for all respondents. We present results on a statistical simulation that assessed whether using re-interviews is a feasible approach in the survey practice of NSIs. Since it is cost-inefficient to re-interview all respondents in the design, only a sample of respondents is re-interviewed. However, the size of this sample will determine the accuracy by which the measurement model can be estimated and thus it impacts the efficiency of adjusted estimators. However, the size of the sample will strongly determine costs of the re-interview. Clearly, there is a trade-off in costs, re-interview sample size, and the efficiency of the adjusted estimators. Further factors are the strength of selectivity between modes and the size of measurement effects. We present results and discuss implications for the viability of the approach at NSIs.

Applying TSE Framework to Comparative Surveys II
Session Chair: Brad Edwards

New Ideas in Sampling for Surveys in the Developing World
Jill Dever[1], Stephanie Eckman[2], Kristen Himelein[3]
RTI International[1], Institute for Employment Reserach[2], World Bank[3]

Many developed countries have high quality census data and/or population registers that can be used to build sampling frames for surveys. In other countries, however, census data is out of date or traditional sampling methods are impractical or dangerous. Multinational surveys very often include one or more countries where traditional sample designs do not work. Problems may occur at the first design stage, in which clusters are selected with probability proportional to size, due to out of date or unavailable census data. Problems can also arise at later design stages, such as persons or households selected within the clusters, because no register data are available or listing households within the cluster is not feasible. This chapter describes the options that are available to samplers in such situations. Techniques to be discussed include: random geographic cluster sampling and nighttime lights at the first stage; and reverse geocoding, random walk, respondent-driven sampling, and quota sampling at a subsequent stage. For each method, we describe the statistical properties and note the pros and cons. Throughout, we suggest the best sampling techniques as ones that minimize interviewer discretion and contain built-in opportunities for verification of interviewer performance.

Improving Cross-National/Cultural Comparability Using the Total Survey Error Paradigm
Tom Smith[1]
NORC at the University of Chicago[1]

Total survey error (TSE) is a very valuable paradigm for describing and improving surveys, but it can be improved. One key limitation is that TSE was formulated to apply to a single, standalone survey. Yet most survey research combines and compares surveys. TSE can be extended to cover these multi-survey utilizations. TSE needs to be thought of as heavily involving the interaction of error components and the concept of comparison error should be used to extend TSE to cover multiple-surveys including trend analysis, comparative studies, and longitudinal panels. This extension of TSE will greatly improve the design of multi-surveys in general and of comparative (i.e. cross-national/cross-cultural) surveys in particular. Likewise, using TSE can greatly advance the analysis of comparative data by using it to assess and adjust for difference in the error structure across surveys. A comprehensive TSE typology should be used whenever comparative studies are designed and also whenever secondary analysis of comparative studies is carried out. In particular strict application of the TSE paradigm can help to achieve the goal of functional equivalence cross-nationally/culturally. Minimizing TSE is an important goal in survey research in general and is especially valuable for comparative survey research and the TSE paradigm should be used as both an applied application and a research agenda to achieve that goal. Extensive examples from the ISSP and ESS will be used to demonstrate this approach.

Innovations in Data Collection in Resource-Poor Settings
Beth-Ellen Pennell[1], *Sarah Hughes*[2], *Kristen Cibelli Hibben*[1], *Jennifer Kelley*[1], *Yu-chieh Lin*[1]
Institute for Social Research, University of Michigan[1], *NORC*[2]

The diffusion of affordable technology in developing and transitional countries is facilitating new approaches to data collection and quality control monitoring. This includes the collection of rich paradata, with immediate access to survey and process data (including call records). Self-administered modes such as audio computer assisted self-interview (ACASI) are being used in new contexts as are the use of digital recordings, GPS, digital photography and digital fingerprinting, among other examples. These advances bring both challenges and opportunities. The use of technology can also disrupt traditional organizational structures and models as well as information flow about data production, quality and costs. This presentation will trace these developments with examples from a number of surveys, with an emphasis on studies conducted in developing and transitional countries. The presentation will also look at regional trends as many of the challenges differ by research and technical infrastructure, number of languages, and cultural traditions. Finally, we will look take a look ahead to developments and trends in this area.

Case Studies on Monitoring Interviewers Behaviors in Cross-national and International Surveys
Zeina Mneimneh[1], *Lars Lyberg*[2], *Sharan Sharma*[1]
University of Michigan[1], *Stockholm University*[2]

Quality control is needed during all the phases of a survey lifecycle. Surveys implemented in diverse cultures are no exception. The diversity of space, time, infrastructure, and expertise makes quality control of utmost important. Many cross-cultural surveys are interviewer-administrated and are conducted in less than ideal circumstances where interviewing traditions differ, multiple contractors are hired to implement the work and interviewer workloads vary greatly. Under such conditions, interviewers become an important source of survey error (variance/bias), and close monitoring becomes essential. This presentation addresses interviewers as an important source of error in surveys conducted in diverse cultures. It discusses some of the cutting-edge methods that have been implemented to monitor interviewer behavior and reduce interviewer error in a number of surveys including two panel surveys in India, a national mental health survey in the Kingdom of Saudi Arabia, a cross-cultural attitudinal survey and the European Social Survey. The presentation concludes with a discussion on future directions on quality assurance and control in diverse cultures.

Analytic Error
Session Chair: Clyde Tucker

Invited Presentation: **The Role of Statistical Disclosure Limitation in Total Survey Error**
Alan Karr[1]
NISS[1]

This paper is an argument for the need to include statistical disclosure limitation (SDL) in the total survey error (TSE) paradigm, accompanied by initial evidence of the value of doing so. Almost all publicly released datasets, and many restricted ones, undergo SDL before release: data are altered in order to reduce the risk of disclosing the identities of subjects and the values of sensitive variables. Methods for SDL include deleting subjects or variables, coarsening categorical variables, adding noise to numerical variables, microaggregation, swapping, combinations of methods, and synthesis of some or even all values in the released dataset. Ideally, SDL methods and their parameters are chosen by means of a quantified tradeoff between disclosure risk and data utility, but this is not always so. As do all other sources of survey error, SDL contributes to uncertainty in analyses of the data. Omitting it from the TSE paradigm omits a source of uncertainty, making analyses more difficult, or even impossible. A central point is the SDL is the only form of survey error that is deliberate, and therefore, is controllable. We introduce, illustrate and demonstrate the value of TSE-aware SDL, that is, SDL that is responsive to knowledge about other sources of error. (This is possible since SDL is typically the final step before data are released.) For instance, we demonstrate that for numerical variables whose measurement error distribution is known, adding noise with this distribution maximizes data utility subject to a constraint on disclosure risk. It seems clear that variables with high error frequencies should require less intensive SDL than variables with low error frequencies, and we show with examples that this is so. We also address feedback loops in the "other sources of error"—SDL relationship. In particular, there is a strong link among editing, imputation and SDL, most of which is completely unexplored. We both formulate key questions and begin to answer them. For instance, how much effort should be expended on editing data whose values will be changed during SDL? How can SDL be performed when edit constraints on data records must be respected? From a confidentiality perspective, do edited and imputed values require the same protection as respondent-provided information? Finally, we propose a unified, Bayesian approach to editing, imputation and SD, steps in the data production process that to date have almost always been disjoint. The approach incorporates explicit measures of data utility and disclosure risk, and we present results from a prototype.

Cross-validation for Robust Variance Estimation in the Presence of Several Error Sources
Øyvind Langsrud[1]
Statistics Norway[1]

Cross-validation is a very important tool in many areas of data modeling, especially when prediction is the main objective. Then, subsets of the data are repeatedly left out from model estimation in order to compare predictions against observed values. In linear regression models, residuals from leave-one-out cross-validation can be calculated directly and is known as one type of adjusted residuals. One way of calculating estimates based on survey data is to predict or impute values for the whole population. When this is done by linear regression modeling, identical results can be obtained by linear calibration weighting. Corresponding robust variance estimates can be calculated from the adjusted residuals mentioned above. The present paper will suggest a modified method in order to take into account clustered data and several error sources. Results based on Norwegian survey data will be presented and cross-validation in more general will also be discussed.

Tuesday, September 22, 2015
9:30-11:30 a.m.
Paper Session V

Interactions of Survey Error and Hispanic Ethnicity II
Session Chair: Sunghee Lee

When Race and Hispanic Origin Reporting are Discrepant Across Administrative Records Sources: Exploring Methods to Assign Responses
Sharon Ennis[1], *Sonya Rastogi*[1], *James Noon*[1]
U.S. Census Bureau[1]

The U.S. Census Bureau is researching uses of administrative records in survey and decennial operations in order to reduce costs and respondent burden while preserving data quality. One potential use of administrative records is to utilize the data when race and Hispanic origin responses are missing. When federal and third party administrative records are compiled, race and Hispanic origin responses are not always the same for an individual across different administrative records sources. We explore different sets of business rules used to assign one race and one Hispanic response when these

responses are discrepant across sources. We also describe the characteristics of individuals with matching, non-matching, and missing race and Hispanic origin data across several demographic, household, and contextual variables. The data in this study include federal and third party administrative files used to build and assign race and Hispanic origin data to an administrative records composite. We develop different methods to assign Hispanic origin and race data to the administrative records composite based on demographic information available in the administrative records files. These methods determine which race and Hispanic origin data to assign and from which administrative records file. Once a single response is assigned to the administrative records composite using each method, it is then linked to the 2010 Census data and we evaluate which set of rules result in the highest level of agreement between the administrative records composite and the 2010 Census. We then determine the best method in assigning a response to the administrative records composite. Next, we use multinomial regression models to predict whether a linked Census-administrative record matches on Hispanic origin or race, whether the Hispanic origin or race data do not match, and whether the administrative record does not have any available Hispanic origin or race data. We find that minorities, especially Hispanics, are more likely to have non-matching Hispanic origin and race responses in administrative records than in the 2010 Census. Hispanics are less likely to have missing Hispanic origin data but more likely to have missing race data in administrative records. Non-Hispanic American Indians/Alaska Natives and non-Hispanic Pacific Islanders are more likely to have missing race and Hispanic origin data in administrative records. Younger individuals, renters, single parent households, individuals living in households with two or more people, individuals who responded to the census in the nonresponse follow-up operation (NRFU), and individuals residing in the West are more likely to have non-matching race and Hispanic origin responses. Younger individuals, individuals living in households with two or more people, and NRFU respondents are more likely to have missing race and Hispanic origin responses.

The Impact of Question Format and Respondent Background on Data Quality in a Health Survey

Aaron Maitland[1], David Cantor[1]
Westat[1]

A number of factors may influence the quality of the answers that respondents provide when answering survey questions. The format of the survey questions defines the task around which the respondent answers the questions. Some tasks will be relatively more difficult. A popular format, particularly in visual modes such as self-administered mail questionnaires, is the grid format where a question stem is associated with a list of items. Although these questions are often used to reduce respondent burden by grouping conceptually similar items, the resulting grids may sometimes result in questions that are more visually challenging for respondents. Furthermore, the questions in grids may be more difficult depending on the background of the respondent. Respondents with less cognitive ability or less familiarity with the survey process may have particular problems with grid questions. Preliminary evidence for these hypotheses was found in a recent data collection cycle of the Health Information National Trends Survey (HINTS). Overall, rates of item missing data were significantly higher for questions in a grid format compared to stand-alone survey questions. Questions in grids routinely had missing data rates over five percent, whereas stand-alone questions routinely had missing data rates under one percent. In addition, there were significant differences by education and language. For example, lower educated respondents routinely had missing data rates on grid questions that were two or three times higher than more highly educated respondents. This suggests that respondents with lower levels of education and lower levels of English ability are likely to have more difficulty with grid questions. This paper expands on these findings with respect to grid questions and explores some of the reasons behind why there are differences in data quality by background characteristics.

Consistency of Hispanic Origin Identification in Census 2000, 2010 Census and the American Community Survey

Leticia Fernandez[1], Sonya Rastogi[1], Renuka Bhaskar[1], Sharon Ennis[1]
U.S. Census Bureau[1]

The 2010 Hispanic population in the U.S. was estimated at 50.5 million, and its growth accounted for more than half of the decade's population increase. Along with fertility and migration, some of the growth in the Hispanic population may have resulted from changes in the reporting of Hispanic origin. The question in Census 2000, "Is this person Spanish/Hispanic/Latino?" may be understood as asking about subjective perception of ethnic membership. In contrast, the question in 2010 Census, "Is this person of Hispanic, Latino, or Spanish origin?" may be construed as asking about ancestry, and may result in the inclusion of individuals of Hispanic descent who do not always self identify as Hispanic. In addition, individuals may be inconsistent about reporting as Hispanic, depending on the situation or because their identity has changed. Using a unique large dataset linking individuals who either reported as Hispanic or listed a Hispanic ancestry across their census responses in 2000 and 2010 and the American Community Surveys for 2006-2010, this study

compares the demographic and socioeconomic characteristics of individuals who (a) consistently identified as Hispanic; (b) consistently identified as non-Hispanic; (c) changed between Hispanic and non-Hispanic when answering questions worded differently, and (d) changed between Hispanic and non-Hispanic when answering the exact same question in Census and ACS. One of our objectives is to study whether socioeconomic disparities within the Hispanic population are associated with consistency in reporting as Hispanic. Movement into and out of a Hispanic identity may have implications on social inequalities as research suggests that those who move out tend to have higher socioeconomic status. In addition, we explore racial fluidity pattern by consistency of Hispanic origin reporting, showing that there is an interrelationship between these complex sociological constructs.

Estimating and Adjusting for Cross-Cultural Differences in Acquiescent and Extreme Response Styles

Mingnan Liu[1], Z. Tuba Suzer-Gurtekin[2], Sunghee Lee[2]
SurveyMonkey[1], University of Michigan[2]

While popular in measuring attitudes and opinions in survey research, Likert scales are subject to measurement error, which emerges as response styles. Response styles refer to systematic patterns of response category selection in which respondents show a tendency to choose certain categories more frequently than other categories independent of the question content. The response styles become a larger problem for cross-cultural studies as respondents from different cultural backgrounds are shown to use distinctively different response styles. Differential response styles have been frequently investigated for Hispanics as they are the largest and fastest growing minority group in the U.S. In particular, the differential response styles imply that cross-cultural comparisons and aggregating the data without considering such a tendency may be misleading. In this paper, we specifically focus on examining and adjusting differential response styles for Hispanics and non-Hispanics. In addition, while various statistical methods for examining and adjusting response styles can be found in the literature, they are scattered across multiple disciplines. A lack of comprehensive overview of the statistical methods for examining and adjusting response styles limits the field's ability to fully utilize these existing methods. Focusing on two of the most frequently studied response styles, acquiescent response style (ARS) and extreme response style (ERS), this paper will 1) provide a thorough overview of existing statistical methods developed for estimating the magnitudes of response styles across cultural groups as well as adjusting for style differences in making comparisons and 2) demonstrate actual applications of the statistical methods using cross-cultural data. In particular, we will examine four statistical models as follows: confirmative factor analysis (CFA), latent class factor analysis (LCFA), item response theory models (IRT), and multidimensional unfolding models (MUM). These methods will be applied to one survey data set that includes both Hispanic and non-Hispanic respondents. The results will be compared with respect to the significance and magnitude of the response styles and the cross-cultural comparisons with and without adjustments. To our best knowledge, there is no study systematically examine and compare these different statistical methods for adjusting response styles. All we know is each method has some gain in comparison to an unadjusted result, while it is critical to understand what these methods can and cannot do and what types of data and assumptions are needed for these models to be used.

Considerations of Survey Error in Surveys of Hispanics

Mark Lopez[1], David Dutwin[2]
Pew Research Center[1], SSRS[2]

As the largest and fastest-growing minority population in the United States, Hispanics have become an increasing focus of survey research. The vast body of Hispanic research evidences myriad options regarding sampling, data collection, and weighting, each of which can affect the resulting data about this population. Typical survey designs feature simple random samples (sometimes obtained as part of larger omnibus or general-population surveys), stratified RDD, "top market," and surname designs. In addition, some studies obtain interviews in English only, while others offer both English and Spanish but make choices regarding the use and allocation of bilingual interviewers. Finally, there are a range of considerations in the weighting of Hispanic survey data. Utilizing data from a national omnibus survey, the General Social Survey, and the Pew Hispanic Center National Survey of Latinos, this article explores these three foci: sampling, interviewing language, and weighting. We report on what we find to be best practices and the implications of failing to enact these practices, as measured by bias and variance in survey estimates of Hispanics.

Estimating Total Survey Error

Session Chair: Paul Biemer

Invited Presentation: **ASPIRE – An Approach for Evaluating and Reducing the Total Error in Statistical Products**

Dennis Trewin[1], Paul Beimer[1], Heather Bergdahl[1]
Swinburne University[1]

In 2011, the Ministry of Finance required that Statistics Sweden develop a quality review and improvement system that contained metrics and enabled changes in quality over time to be assessed. The Ministry also required the highest priority areas for improvement to be identified. These requirements led to the development of the system now known as ASPIRE (A System for Product Improvement, Review and Evaluation) which satisfies these requirements. So far, three successful applications of ASPIRE have been undertaken for 10 surveys and other statistical products. This paper will describe ASPIRE and how it satisfies the requirements of the Ministry of Finance and, most importantly, provides valuable information to Statistics Sweden to enable it to target its quality improvement effort. It will be illustrated by case studies from the three applications. The focus of this paper is on the Accuracy dimension of quality but it can easily be extended to the other quality dimensions and some pilot studies have been undertaken to demonstrate this. ASPIRE requires a separate assessment of Accuracy for each of five criteria (knowledge of risks, communication to users and providers, available expertise, compliance with standards or best practice, and planning/mitigation of risks) and the relevant sources of error for the type of product. In deriving the overall product rating, the intrinsic risks from the various sources of error are taken into account. The assessment process will be described with suggestions on how it might be extended to other providers of statistical products. We will pay special attention to the evaluation of the National Accounts (particularly Annual and Quarterly GDP) as well as the business collections that are key inputs to the National Accounts. Our approach is innovative and provides an objective approach to evaluating the Accuracy of the National Accounts and identifying highest priority areas for improvement. In particular, a specific error structure has been designed for the National Accounts to assist with the evaluation. Research Triangle Institute, (2) Former Australian Statistician, (3) Statistics Sweden.

Invited Presentation: **Total Survey Error Assessment for Socio-Demographic Subgroups in the 2012 National Immunization Survey**

Benjamin Skalland[2], Vicki Pineau[2], Wei Zeng[2], Kirk Wolter[1], Meena Khare[3], David Yankey[3], Phil Smith[2]
NORC University of Chicago[1], Centers for Disease Control and Prevention[3]

Total survey error (TSE) modeling provides a framework for evaluation of sampling and nonsampling errors in statistics for young children aged 19-35 months and teens aged 13-17 years produced from the National Immunization Survey (NIS), a vaccination surveillance program conducted on an ongoing basis by the U.S. Centers for Disease Control and Prevention. The TSE modeling developed for the NIS is conducted in three broad steps for an NIS estimator under study: (1) specify a distribution function for each component of error in the survey assuming independence of each component error, (2) derive estimates of these component distributions from the best sources available, and (3) apply a Monte Carlo simulation approach to combine all components of error into a TSE distribution for the survey estimator. The mean of the TSE distribution provides an estimate of the total bias in the survey estimator. We have previously assessed TSE since 2006 for several key vaccination coverage rate estimators at the national level for the total populations of young children and teens. In this presentation, we use 2012 NIS data to assess TSE in estimated vaccination coverage rates for selected socio-demographic subgroups defined by metropolitan status, race, Hispanic ethnicity, and income. We compare estimates of bias derived from the TSE analysis for total population and for the socio-demographic subgroups.

Systemic and Aggregate Components of Total Error in Sample Surveys and Administrative Record Systems

John Eltinge[1]
BLS[1]

In the development and use of total survey error models, it can be important to distinguish between error sources that are aggregate and systemic, respectively. An aggregate error source arises from the combined effects of a large number of (approximately) independent random events. Examples include reporting errors or item nonresponse patterns that arise from events that take place in an individual sample household or establishment. With some exceptions, most of the previous literature on total survey error has been based on aggregate error models. On the other hand, a systemic error source arises from a single event, or small set of events, that can affect the quality of a large number of responses. Examples include omission of subpopulations from frames; irregularities in the definition of strata or primary sample units; mistakes in the computation of selection probabilities; errors in programming of CAPI or CATI instruments; and

training or management problems that degrade the performance of groups of interviewers. This paper explores four general issues that arise with systemic errors. First, we review possible models for some classes of systemic errors, and note that analyses of these models can be especially challenging because some standard large-sample statistical properties may not apply. Consequently, it is of special interest to adapt methods developed previously for the reliability analysis of large and complex systems. This includes methods arising both from frequentist approaches and from elicitation of prior distributions and utility functions for Bayesian analyses. The resulting models lead to characterization of the impact of systemic errors on both conditional and unconditional bias and variance properties of standard estimators. Second, the abovementioned challenges also lead to sensitivity analyses that assess the potential impact of systemic errors in a given survey. These sensitivity analyses are based on extensions of models developed originally for aggregate error sources under a total survey error approach. In addition, these analyses provide a relatively simple way to identify systemic error sources that warrant the more elaborate modeling work described above. Third, we provide a framework for characterization of the potential impact of a systemic error. This framework accounts for the magnitude of the effect that the error has on the statistical properties of an estimator; the duration of that effect; and the resources required to identify and address these effects. This in turn suggests a two-step approach to management of systemic errors: efforts to prevent the occurrence of systemic errors where feasible; and design of survey processes to be relatively robust against the impact of such errors when they occur. Finally, we consider related issues with systemic errors that arise with data obtained through non-survey sources, e.g., administrative records, commercial transaction information and other forms of organic or "big" data. Special attention is directed toward systemic errors that arise from the administrative, commercial or social processes that lead to the production and capture of these non-survey data.

Working Toward an Estimator for Total Survey Margin of Error

Natalie Jackson[1]
Huffington Post/Pollster.com[1]

Researchers know the traditional "margin of error" reported with surveys only accounts for sampling error, which presents an incomplete estimate of survey error. The math used to calculate MoE assumes true random sampling and that all members of the population have a known probability of being selected, but in practice, virtually no single sample frame for a population fits these strict mathematical criteria. Indeed, when statisticians introduced the concept of confidence intervals, the example for a random sample was to start at the center of a perfect circle and randomly select in which direction to draw a radius out to the edge of the shape (Neyman 1937). Few samples of humans could ever hope to achieve that kind of true randomness and equal probability. Thus, the margin of error is, at best, a highly flawed estimate of survey error. In this paper, I will discuss my efforts to develop a more complete survey error estimator. This estimator needs to be flexible enough to work with mixed mode or multi-frame surveys and incorporate far more information about the survey process than the simple margin of error. The estimator I'm developing will do the following, with the ability to add more steps: 1: Account for differences between the population and the sample frame(s) to estimate of coverage error 2: Account for differences between the population and the sample to estimate sampling error 3: Account for differences between the sample frame(s) and the sample to estimate nonresponse error 4: Account for weights (or model-based) adjustments on sample: post-survey error. The user would need to input information about their sample and the sample frame, enter the response proportions for the survey item of highest interest, and a statistical routine will run to do the four estimates listed above. Theoretically, estimates of error on different survey items would be close to the same, so only one item would need to be calculated (this will be tested). Then, the routine would run simulations on a beta or dirichlet distribution (depending on whether the survey item is binary or multi-category) using the error calculations as parameters in the estimation, with the reported proportions as the distribution means. The routine would then return values for the ranges of the simulated distributions that would represent probabilistic ranges for the population parameter given that sample. The number for 95 percent of the simulated distribution could then be translated into a margin of total survey error.

Comparing the Mean Square Error Between Alternative Survey Design Procedures

Gary Shapiro[1], *Keith Rust*[2]
Statistics Without Borders[1], *Westat*[2]

In comparing a less expensive but biased survey design to a more expensive design, it is common to assume that the more expensive design is completely unbiased, This is an unrealistic assumption that may lead to a wrong decision. This paper compares results for alternative values for the bias of the more expensive design, as well as for alternative added bias for the less expensive procedure and alternative variances for the two designs.Graphs are given which show the relationship in root mean square error of the two design choices. In general, when the sample size is large and one is not dealing with

particularly small domains, the more expensive design is likely to be the best choice, even if the standard error is much lower for the less expensive design. Thus, assuming complete unbiasedness for the expensive design can sometimes lead to an incorrect decision.

Measurement Error

Session Chair: James Wagner

Comparing the Quality of 2010 Census Proxy Responses with Administrative Records

Mary Mulry[1], Andrew D. Keller[1]
U.S. Census Bureau[1]

Currently the U.S. Census Bureau is conducting research on ways to use administrative records to reduce the cost and improve the quality of the 2020 Census Nonresponse Follow up (NRFU) at addresses that do not self-respond electronically or by mail. In previous censuses, when a NRFU enumerator was unable to contact residents at an address, he/she found a knowledgeable person, such as a neighbor or apartment manager, who could provide the census information for the residents, called a proxy response. The Census Bureau's recent advances in merging federal and third-party databases raise the question: Are proxy responses for NRFU addresses more accurate than the administrative records available for the housing unit? Our study attempts to answer this question by comparing the quality of proxy responses and the administrative records for those housing units in the same timeframe using the results of 2010 Census Coverage Measurement (CCM). The assessment of the quality of the proxy responses and the administrative records in the CCM sample of block clusters takes advantage of the extensive fieldwork, processing, and clerical matching conducted for the CCM.

Measuring Financial Literacy in a Large-Scale General Survey

Jonas Beste[1], Arne Bethmann[1]
Institute for Employment Research[1]

In a variety of research questions, e. g. the risk of poverty, the relationship between income and living conditions or satisfaction with personal income, the ability of a person to use knowledge and skills to effectively manage financial resources is a relevant factor. A person who exhibits a high degree of this ability is likely to have a lower risk of poverty, a more efficient conversion of income into living standard and a higher satisfaction with his income. This ability is described by the concept of financial literacy. Although, it is substantial for many research projects, financial literacy is seldom measured directly in social surveys. Often the individual education is used as a proxy. Previous research has shown that education is a weak proxy for financial literacy. To take the financial literacy of a person into account investigating this kind of research questions, we developed an eight item Likert scale which ran in 2014 on the 8th wave of the German panel study `Labour Market and Social Security' (PASS). PASS is a longitudinal data set for Germany that focuses on welfare receipt and labour market participation but is also usable to give evidence about a variety of subjects for the German general population. The question battery covers three theoretically dimensions: planning of finances, handling current finances and financial pressure. In this presentation we introduce our measurement instrument for financial literacy and show same descriptive results. To verify that the same construct is measured across different groups, we perform tests of measurement invariance. Here, measurement invariance is tested in the framework of multi-group confirmatory factor analysis (CFA). We focus on differences between person with low and high income as well as person with low and high education. It could be assumed that those groups differ in the makeup of the underlying construct so the mean-values could not be simply compared to each other.

The Influence of an Up-Front Experiment on Respondents' Recording Behaviour in Payment Diaries: Evidence from Germany

Tobias Schmidt[1], Susann Kuehn[1]
Deutsche Bundesbank[1]

In this paper we analyse the recording behaviour of German consumers in a one week diary on their point-of-sales expenditures. We are particularly interested in the effect of an experiment, eliciting respondents' risk preferences, on their recording behaviour. The experiment is run shortly before the consumers start to fill in the diary. In the experiment the consumers have the choice between receiving a sure payment of 10 euro and participating in a game. If they opt for playing the game they roll a die and either win 20 euro if it shows 4, 5, 6 or nothing if it shows 1, 2, or 3. We ask whether

respondents' recoding behaviour differs depending on whether individuals who do play lose or win. We argue that winners may attach a more positive feeling to the survey than losers and therefore exhibit more commitment to the diary, e.g. by reporting better quality data. Beyond providing evidence on the effect of conducting up-front experiments in representative surveys our results also contribute to the literature on participation incentives. For participants who roll the die the experiment can be seen as an incentive experiment in which consumers are randomly assigned an incentive of zero or 20 euro. Several measures of data quality and recording behaviour can be used in the analysis, including the number and type of transactions recorded, the number of missing values and the pattern of recorded transactions between days. Our results indicate that the outcome of the game has an impact on the quantity of transactions recorded, but does not affect the quality of information recorded and measures like the cash share.

Room for Error: Rating Scale Inconsistencies and Solutions

Joseph Goldman[1]
The Gallup Organization[1]

Every survey methodologist seeks to find the best way to measure nuances in public opinion. Rating scales, one of the most commonly used techniques for measuring varying intensities of an opinion, appear to offer an elegant solution to categorizing a respondent's opinion. When used in domestic surveys, many polling organizations and research firms have found them to be effective. International comparisons offer a different level of efficacy. The European Values Study (EVS), World Values Survey (WVS), and many other international research tools offer a number of rating scales (including many Likert items) to compare public attitudes across countries. However, comparing these statistics internationally proves to be problematic. For example, a number of EU member states with higher tax compliance statistics (Schneider 2011) appear have lower tax morale (Goldman 2013) according to the EVS. When asked a 10-point rating scale on the subject of morality of tax evasion, various abnormalities occur, including abnormal frequencies of particular numbers in countries along with a geographically disproportionate number of respondents giving "extreme" answers. The nature of such the questions results in a distribution of responses that is unhelpful and misleading to major consumers of data for this and other questions. The improper use of these statistics by researchers, especially in the fields of economics and econometrics, leads to potentially inaccurate research outcomes. Alm and Torgler (2004) utilize a methodology that recalculates cultural variables, apparently for statistical convenience. While this treatment does simplify data analysis and remove apparent outliers, it does not fix the root of the problem, which is based in unclear answer options. In addition, statistical manipulation of the individual-level values (country-level means, etc.) manipulate the very meaning of the response. Other options, such as dichotomous questions, avoid many of these concerns. Many recently introduced international surveys already use simplified questions to enable respondents to choose meaningful responses. Along with rigorous pre-testing, more user-friendly data can be easily analyzed and interpreted by researchers from other fields to make valuable policy decisions.

The Multi-Trait Multi-Error Approach to Estimating Measurement Error

Alexandru Cernat[1], *Daniel Oberski*[2]
University of Essex[1], *Tilburg Unviersity*[2]

Measurement error is a pervasive issue in surveys. One of the most common approaches used to measure and correct for systematic errors in this context is the Multi-Trait Multi-Method approach. Thus, it is possible to separate method, random error and "true" score using an experimental design that combines multiple traits (i.e. questions) with multiple methods (i.e. answer scales). As with other statistical approaches that tackle measurement error the results of this model are biased if any other types of systematic error (such as social desirability) are present. In this paper we present an extension of this model, which we name Multi-Trait Multi-Error, that manipulates multiple characteristics of the question format using a within factorial design. Thus, it is be possible to estimate simultaneously: social desirability, acquiescence, method, random error and "true" score. We will illustrate how to implement the design and show initial results using measures of attitudes towards immigration in the 7th wave of the Understanding Society Innovation Sample.

Does Big Data Mean Lower Quality Data?

Session Chair: Alan Karr

Relationships Between Data Quality and Confidentiality

Jerome Reiter[1]
Duke University[1]

Most agencies redact data before sharing them with the public, for example by adding noise to values, swapping variables across records, or suppressing values. These data redactions necessarily have implications for data quality. In this talk, I offer thoughts on how perturbations caused by disclosure limitation techniques connect with data quality.

Total Error and the Analysis of Big Data: Why Size Doesn't Matter

Paul Biemer[1]
RTI International and University of North Carolina[1]

Big Data involve massive amounts of very high-dimensional and unstructured data that bring both new opportunities and new challenges to the data analyst. Some of the errors that plague Big Data are well-known. As they are created, Big Data are often selective, incomplete and erroneous. However, new errors can be introduced downstream as the data are cleaned, integrated, transformed, and analyzed. The data munging (or data wrangling) steps that often comprise 50-80% of the work involved in getting a dataset ready for analysis can add both variable and systematic errors to the data, resulting in unreliability, invalidity, and biased inference. This paper considers the 'total error' associated with Big Data and demonstrates some ways these errors can lead to false discoveries, invalid inferences and poor business decisions. We present a total error model for Big Data that enumerates the major sources of error and show how these errors can affect Big Data analytics in ways that may only be exacerbated by their large size. Some statistical approaches for minimizing these risks borrowed from classical statistics will be considered.

Dealing With Nonsampling Error in PIAAC

Session Chair: Tom Krenzke

Associations Between Interviewer Insights and Proficiency Scores

Michael Lemay, Valerie Hsu[1]*, Richard Sigman*[1]*, Tom Krenzke*[1]
Westat[1]

The 2012 Programme for the International Assessment of Adult Competencies (PIAAC) involved collecting interviewer observations regarding the circumstances under which the assessment took place. The relationship between interviewer observations and the key survey outcomes (proficiency measures in literacy, numeracy and problem solving) could shed light on possible data collection protocol changes. This presentation discusses results from an analysis that aims to associate interviewer insights with resulting proficiency scores. We found that the respondents' survey-taking environment as well as respondents' behavior and characteristics were related to their assessment outcome. Interviewer observed economic status and whether or not the respondent asked for clarification while undertaking the interview were the strongest predictors of proficiency scores after considering all other interviewer observations. Interviewer-observed economic status, which was also collected for nonrespondents through the non-interview report form, could be useful as a household-level variable for both respondents and nonrespondents for reducing nonresponse bias during weighting since the observed economic status and proficiency measures were highly correlated. Interviewer insights empower survey practitioners to develop better data collection protocols in order to minimize measurement error and to identify effective weighting variables in order to reduce nonresponse bias.

PIAAC Japan Nonresponse Bias Analysis

Takahiro Tsuchiya[1]
The Institute of Statistical Mathematics[1]

We examine the potential bias of the PIAAC Japan results by comparing demographic characteristics between respondents and nonrespondents as well as comparing proficiency scores between easy-to-access and hard-to-access respondents. In terms of demographic characteristics, the response rates differed at most six percentage points among gender and age, though the response rates were more than 15 percentage points lower in urban and highly educated areas, where mean proficiency is high. Considering merely the area level response rates, the mean proficiency estimate without a weighting adjustment is supposed to be biased downward. Actually, the proficiency estimates using nonresponse adjusted weights becomes larger than the base-weighted estimates. Further investigation into employed and highly educated people, whose proficiency is likely to be high, didn't seem to refuse the survey. This view is supported by the fact that the proportion of the well-educated is higher in respondents than in population. Hence, the mean proficiency estimates using weights that were calibrated to population after nonresponse adjustment are smaller than the base-weighted estimates. In terms of degree of easy-to-access, we examined five variables, where two were related to the possibility of in-home, and the other three were related to cooperative attitudes of respondents toward survey. These two factors were mainly

considered because the major reasons of nonresponse were absence and refusal. In order to evaluate how often respondents are absent, we counted the number of interviewer's visits before completing the interview, and we directly asked respondents "Do you usually stay at home or stay out?" Although the mean proficiency shows little differences among the number of visits, people who typically stay out according to self-report show high proficiency in almost all the demographic characteristics. As for cooperative attitudes, we directly asked respondents "If you were asked to participate in a similar survey again, what would you do?", and we asked interviewers to evaluate how much the respondent was cooperative and interested in the survey. Using either variable, the more cooperative respondents showed higher proficiency in every demographic characteristic. In short, the absent nonrespondents might have higher proficiency while the refused nonrespondents might have lower proficiency than respondents. Considering the fact that about three quarters of nonrespondents refused the survey, mean proficiency of nonrespondents could be lower than that of respondents.

Treatment of Inaccessible Sampling Units in an International Survey of Adult Competencies

Leyla Mohadjer[1]
Westat[1]

Sponsored by the Organisation for Economic Co-operation and Development (OECD), the Programme for International Assessment of Adult Competencies (PIAAC) offers continuing global assessment of adult skills in multiple cycles. The first cycle of PIAAC consists of three rounds with participation from about 40 countries. The data collection for the first round of cycle I of PIAAC was conducted in 2012, and the second round was conducted in 2014. Another group of countries are starting to prepare for the Round 3 of cycle I data collection, planned for 2017. Similar to other international comparative studies, the goal of PIAAC is to produce data that make inferences and comparisons across national populations on the basis of survey samples selected from the same target population in each country (i.e., non-institutionalized adults 16 to 65 years old residing in the country). To achieve this goal, all participating countries are required to follow consistent guidelines covering all aspects of the study to facilitate valid comparisons of survey results internationally. This paper focuses on the requirement for ensuring consistency in coverage of the target population across countries. The paper discusses the challenges in achieving such consistency given the differences in the types and quality of sampling frames across countries, and then focuses on challenges in using registries as sampling frames in international surveys. More specifically, the paper will discuss 1) the challenges in using registries with incorrect/out-dated information about the sampling units, 2) how sampling units with incorrect information are treated in national surveys, and 3) offer alternative approaches for dealing with such cases in international surveys. The discussions and conclusions are based on lessons learned from the first round of PIAAC. We describe the difficulties PIAAC countries had in locating sampled persons selected from registries with incorrect addresses (referred to as inaccessible persons in registry samples). We then present alternative approaches countries had and the options they chose for dealing with inaccessible persons in the PIAAC sample.

Is Self-reported Health Status at the Time of Interview Associated with Respondents' Performance on the Assessment?

Lin Li[1], *Tom Krenzke*[1], *Martha Rozsi*[1], *Leyla Mohadjer*[1]
Westat[1]

This paper explores whether or not how well we feel at the time of interview is associated with results from a test to measure proficiency levels in literacy, numeracy and problem solving. Using the data from the 2012 Programme for the International Assessment of Adult Competencies (PIAAC), we looked at the relation between self-reported health status and proficiency scores across countries. Although previous research has shown that age is strongly related to proficiency (older are less proficient), the relationship between health status and proficiency is less clear. If associated, then measurement error would exist in the results of the PIAAC assessment and potential for change in data collection protocols could be justified. Therefore, the proficiency scores' association with self-reported health, age, and other key demographic characteristics was thoroughly investigated. PIAAC collects self-reported health status instead of physical health, therefore, to further investigate the quality and meaning of self-reported health status, we reviewed the relationship between physical health and self-reported health from National Health and Nutrition Examination Survey (NHANES) data.

Discussant: Lars Lyberg, Stockhol University

Uses of Geographic Information Systems Tools in Survey Data Collection & Analysis

Session Chair: Ned English

Virtual Canvassing: In-Office Methods for Validating the Census Bureau's Address List for the 2020 Census

Michael Ratcliffe[1], Shonin Anacker[1], April Avnayim[1], Christopher Henrie[1], Tiernan Erickson[1], Dakota Schuck[1]
U.S. Census Bureau[1]

Assuring an accurate and complete address list for the United States and Puerto Rico is a critical step in the Census Bureau's planning and conducting of each decennial census. For the 2010 Census, the Census Bureau conducted a full address canvassing operation, with field workers traversing nearly every road in the nation to verify and update addresses in its Master Address File (MAF). Analysis of address and land use/land cover data from a variety of sources indicate that modeling and data-driven decision-making can substitute for expensive in-field operations. As a result, for the 2020 Census, the Census Bureau plans to implement in-office methodologies to validate the accuracy and completeness of the address list for most of the housing units in the nation. In this presentation, I report on the variety of address, land use/land cover, and imagery-based data under consideration for use in evaluating and validating the completeness of information for addresses in the MAF for small geographic areas and identifying areas in which in-office canvassing is appropriate and viable. In addition, I discuss the methodologies and tools for managing data for over 130 million addresses and 11.1 million census blocks.

Using GIS to Understand Error Sources in a Web Survey

Ned English[1], Michael Stern[1], Ipek Bilgen[1], Ilana Ventura[1]
NORC at the University of Chicago[1]

The web mode has been given considerable attention in recent years as a potential alternative to random-digit dial telephone surveys, due to the potential for cost efficiencies as well as improved coverage and lower rates of non-response. We would expect web surveys, however, to tend to be more successful in certain geographic places than others, as a function of internet access and proficiency. As a result, there exist challenges to probabilistically recruiting general population households to a web survey from a random sample of addresses. Our paper presents results from three studies where households were contacted both via physical mail and emails that were matched to selected addresses in a diverse state with varying geographic and socio-demographic environments. We use spatial modeling within geographic information systems (GIS) to understand how the kinds of people who respond to web surveys at as recruited by either mode of contact may differ from the population at-large. In so doing, we consider not only what types of households respond to a given mode or contact method, but also the relationship between geography and coverage or non-response bias by either contact method. Our research is useful for survey methodologists who are considering implementing a web survey and want to understand and visualize the potential for coverage and non-response error.

The Role of Geographic Information in Minimizing TSE for a Large-scale Natural Resource Survey

Sarah Nusser[1], Emily Berg[1], Alan Dotts[1], Zhengyuan Zhu[1]
Iowa State University[1]

The National Resources Inventory (NRI) monitors status and trends in land cover and land use over time. As a land-based survey, the NRI makes use of geospatial information at various stages of the survey process, including collection, editing and error checking, and estimation. A location certification process, involving orthorectification of digital images, improves the precision and accuracy of location information. This ensures that observed changes in land characteristics are not confounded with modest shifts in the location of the image. Through the use of a web application, data collectors for the Conservation Effects Assessment Project delineate agricultural fields on images of sampled primary sampling units. In NRI estimation, Geospatial data on federal and large water areas provide auxiliary information in construction of the unit-level data set. The primary source for external information on large water is the National Hydrography Database (NHD), and information on federal areas is obtained from numerous agencies. Although one reason to incorporate auxiliary data on federal and large water areas is to reduce sampling errors in estimators through calibration, the external information also helps reduce non-sampling errors. In particular, the auxiliary data can identify measurement errors in NRI collected data and improve consistency with external data sources.

Using Geospatial Analysis to Inform Household Survey Design Decisions and Harvest Sample Efficiencies

Rosemary Byrne[1], Aliza Kwiat[1]
US Census Bureau[1]

The analysis of spatial relationships is a valuable tool in understanding what drives policies or natural phenomena to function differently in different places. The US Census Bureau produces many excellent visualizations and info-graphics displaying social and economic trends. In our work creating efficient and effective sample designs for several of our household surveys, we wanted to demonstrate the benefit of applying this type of space-based approach to our day-to-day functions. We partnered with geographers to bring our statistical knowledge to bear on how best to communicate the underlying truths in our data. In this paper, we discuss several geospatial applications we developed. These range from efforts to target resources where they are most needed; to creating a platform for evaluation of data or processes in collaboration with others. We will show that using spatial analysis and visualizations allows us to convey complicated aspects of our work in a succinct and clear manner.

The Value of Self-reported Frequently Visited Addresses in GPS Assisted Travel Surveys
Timothy Michalowski[1], Dara Seidl[1], Rena Peña[1]
Abt SRBI[1]

Abt SRBI has extensive experience in the application of GPS technology for Household Travel Surveys (HTS) through deployment of 10,000+ GPS devices for various projects throughout the USA. Households participating in GPS travel surveys are deployed small personal GPS loggers for each eligible household member. Participants are instructed to carry the GPS loggers for all travel for 3 to 7 days. The device records GPS data every 1 second, typically resulting in 10,000+ GPS data points per person per day. The GPS travel data demonstrate that respondents typically underreport trip data in travel diaries either due to respondent burden of diary reporting, or respondent misunderstanding of what constitutes a "trip". Short and non-vehicle trips are often underreported, particularly in urban areas. Improved understanding of trip underreporting can significantly enhance the end products of the survey efforts, such as calibrations of regional travel models. As part of the normal industry standard in travel survey recruitment, participants are asked for "frequent travel locations" for all household members. Locations such as home, work, school, and shopping are requested during the recruitment questionnaire. These frequent location data have been used in the survey process to aid in completion of missing self-reported diary trip locations. The use of GPS technology allows for high-precision passive travel data collection as an improvement to reliance on traditional self-reported diary methods and corresponding recruitment efforts. With the high precision and frequency of travel location data provided directly by GPS capture, the role of frequent location capture in recruitment efforts is unclear. This paper examines the value of reported frequent location addresses in recruitment compared with the final GPS travel data products. The frequent locations from travel survey recruitment are geocoded and compared to GPS travel data in Geographic Information Systems (GIS) using spatial statistics. This analysis answers the question, do the frequent locations reported in recruitment enhance GPS travel data? Recent regional household travel surveys completed by Abt SRBI in the USA featuring both GPS and non-GPS populations in each region will be examined. The results of this study inform the level to which the respondent burden of traditional address location reporting is warranted in consideration of rapidly improving GPS technology and analysis capabilities.

Paradata and Responsive Design
Session Chair: Raphael Nishimura

Invited Presentation: **Using Paradata Dashboards to Monitor Interviewer Behavior and Reduce Measurement Error**
Nicole Kirgis[1], Zeina Mneimneh[1], Yan Sun[1], Jay Lin[1], Shonda Kruger Ndiaye[1]
ISR[1]

Over the past several years, there has been an exponential increase in the use of paradata to achieve greater efficiency in data collection and to improve data quality. This has been done at different stages of the survey lifecycle, targeting different sources of errors, and guiding responsive design. In this chapter, the use of paradata in interviewer-administered surveys is explored. It focuses on interviewers as an important source of error and the development and application of paradata dashboards to monitor and control this error. Using specific examples from four case studies, this chapter explores strategies and lessons learned for the use of paradata to monitor and intervene on interviewer behavior. Two of the surveys: The Panel Study of Income Dynamics (PSID) and the National Survey of Family Growth (NSFG) are conducted in the United States. The other two studies are national mental health surveys conducted in China and the Kingdom of Saudi Arabia. Across the different surveys, paradata at the interviewer level, including keystroke data and

time stamps from the computer-assisted personal interviews (CAPI) are used to create multiple indicators of data quality and monitor interviewer behavior at various points during data collection. These indicators include the average time spent on survey questions, the number of questions asked, the pauses taken by interviewers during the interview, the amount of backing up in the interview, and the handling of error checks while administering the survey. Other types of paradata used include the maximum recurrence of similar survey responses, verification results, and item missing data rates. Paradata can also be used to address various constraints and challenges of interviewer-administered data collection. For example, in a study the size of the mental health survey conducted in China, using paradata to focus and prioritize resources is paramount to data quality monitoring at the interviewer level. This chapter also discusses how paradata was used to target the review of recorded interviews and to schedule verification interviews. To allow for timely intervention, automation of the delivery and display of quality indicators to field managers is essential. Different tools can be used for this purpose. This chapter will present some of the tools used, including paradata dashboards and the OLAP (Online Analytical Processing) Cube. Making these tools available at the beginning of data collection is extremely important for timely intervention on interviewer behavior in order to reduce interviewer-related errors in surveys and improve data quality. References: Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G., and Kruger-Ndiaye, S. (2012). Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection. Journal of Official Statistics, 28(4), 477-499. Kirgis, N. and Lepkowski, J.M. (2012). Design and Management Strategies for Paradata Driven Responsive Design: Illustrations from the 2006-2010 National Survey of Family Growth. Chapter 6 in Improving Surveys with Paradata: Making Use of Survey Process Information, Frauke Kreuter, editor. New York: J.W. Wiley and Sons.

Invited Presentation: **Measurement Error in Survey Operations Management: Detection, Quantification, Visualization, and Reduction**

Brad Edwards[1]
Westat[1]

In the total survey error paradigm, nonsampling errors and their relationship to cost have been very difficult to quantify, especially in real time. Recent advances in paradata and Big Data processing and analysis offer an opportunity to address this problem. For example, CARI data selected with known probabilities from a pretest could be used to produce estimates of questionnaire (specification) error, to make improvements to address the design problems, and to monitor error levels after changes are implemented in the main data collection phase. (Hicks, Edwards, Tourangeau, et al. 2010). Latency data from web surveys could be used in a similar fashion. GPS data from face-to-face surveys can detect falsification as it happens, thereby improving quality and saving costs that could be directed elsewhere. The quality improvement could be estimated by comparing the level of falsification detected with GPS compared to the level detected by more traditional methods (e.g., mail return forms, telephone and in-person re-interviews, and CARI coding). Savings from this innovation over the life of a survey's data collection cycle could be estimated by comparing the GPS costs with the costs of detecting and remediating falsifiers using traditional methods. This paper summarizes some recent developments and studies in CARI, GPS, mobile technology, and call record data, informed by the TSE paradigm. ·
For CARI, we plan to implement and evaluate the suggestion by Hicks and her co-authors in a CAPI/ACASI survey of women and sexual abuse, with data collection scheduled late this summer through the winter of 2015. CARI coding will be accomplished with a dashboard that displays a host of variables associated with the question, the interview, and the interviewer (e.g., timing data, interview length and time of day, number of instances the question was asked by the interviewer, etc.) This may become the first instance of quantifying specification error in a sample of question administrations.

· For GPS, we plan to report on a large population study's experience running GPS falsification-detection techniques alongside more traditional methods. Implementation includes visualization of the GPS data in a supervisor dashboard and daily reports on interviews that have a high probability they were falsified in the past 24 hours. Comparing the two methods will provide a measure of false positives and false negatives. We will also analyze costs of the two methods.

· For mobile devices, we will evaluate two small pretests and the first production implementation of several applications on a web-connected smart phone deployed in a CAPI/ACASI survey. Time per completed interview is a common cost metric on surveys that use data collectors, and is the cost driver for these surveys' budgets. In one application the data collector will track her time use throughout the day. Quality of call record data is notoriously bad, yet these data are increasingly used to monitor data collector efficiency and attempts to obtain response. The second application will be a simplified way to track attempt outcomes (not home, refusal, etc.) on a mobile device as they occur, integrated with the field management system and the data collection software. In this way, the supervisor can use the outcomes in real time to text or send an instant message to the data collector, to provide advice or to troubleshoot. The third "application" will be GPS coordinates throughout the data collector's work day, linked with the activity data. These

data will be used to create routing maps for supervisor review with the data collector. All three of these studies – CARI, GPS, and mobile devices -- will address issues in quantifying nonsampling errors and costs, using visualization and process quality monitoring techniques.

Measurement Error in American Community Survey Paradata and 2014 Redesign of the Contact History Instrument

Matthew Virgile[1], Rachael Walsh[1]
U.S. Census Bureau[1]

In recent years, many studies have addressed the potential of using paradata to reduce total survey error. There is less research, however, on the quality of paradata itself and how it may be prone to measurement error. Missing, incomplete, or inaccurate paradata may bias estimates of total survey error. Thus, paradata quality is crucial. In this research, we analyze paradata from the U.S. Census Bureau's American Community Survey (ACS) to see if modifications to the Contact History Instrument (CHI) reduce measurement error and improve paradata quality. The CHI is a data collection instrument used since 2004 by Census interviewers to collect data on contact attempts in demographic surveys. Historically, CHI data have contained many empty records, including prefilled information about sample units but empty fields pertaining to the contact attempt. Since its inception, the CHI has provided a discretionary "just looking" option allowing interviewers to exit the CHI without entering information, generating most of these empty records. It was unclear, however, whether these empty records were due to interviewers bypassing the CHI after making a contact attempt or performing some action other than a contact attempt that the CHI did not capture. In January 2014, the Census Bureau launched a redesign of the CHI with substantial changes to many of the questions. This included removing the "just looking" option, enabling interviewers to specify explicitly if they were not attempting contact and what actions they were performing instead, such as reviewing case information or locating activities. At the interviewers' request, an additional change differentiates between incoming and outgoing telephone contact attempts. The answer categories recording respondent concerns and contact strategies attempted by the interviewer also underwent changes to both reduce the number of options as well as provide clarification. This research explores the impact of the 2014 redesign on the quality of the paradata produced by the CHI. Using data from the 2013 and 2014 ACS, we use descriptive statistics to compare changes in the distributions of several variables. Our results show a substantial reduction in empty records from 24 percent in 2013 to less than 2 percent in 2014, suggesting that the new instrument reduces measurement error and improves paradata quality. This reduction is due to interviewers selecting the new "Not Attempting Contact" option, which further enhances paradata quality by providing information on the level of effort expended in addition to making contact attempts. The new options for this path provide a good representation of other interviewer activities, given our results show the "Other" category is only selected 5 percent of the time. By comparison, the modifications to the answer categories minimally impacted these variables. We observe a 1 percent increase in the selection of the "Other" category for respondent concerns and 3 percent for contact strategies attempted. We also examine the redesign impact through comparisons of number, type, and outcome of contact attempts between years. Finally, we include factor analyses on the new respondent concerns categories to see whether constructs captured in prior research are still being measured.

Tuesday, September 22, 2015
1:30-3:30 p.m.
Paper Session VI

Nonprobability Sampling Methods from a Total Survey Error Perspective
Session Chair: Raphael Nishmura

Probabilistic Sampling with Quotas: A New Look at an Old Method

Neale EL-DASH[1]
Sleek Data[1]

In this paper I examine the sampling and inference methodology of polls in Brazil. The most used sampling designs is Probability Sampling with Quotas (PSQ).This sampling scheme has two stages, where in the 1st stage clusters are selected, usually census tracts, and in the second stage, the selection of the actual respondents is done in a non-probabilistic fashion, using quotas.This sampling design is criticized by academics because it doesn't allow the inclusion probabilities for all respondents to be calculated, and therefore it is not possible to obtain the estimates usually recommended for the quantities of interest. The aim of this paper is to present a model-based justification for PSQ and compare it, from the

point of view of design-based inference, with an equivalent fully probabilistic design. The use of the response homogeneity group model (RHG) to explicitly model the probabilities of individual response allows the use of the usual estimators.The same model for the probabilities of response allows calculation of the inclusion probabilities for the case of probabilistic sampling, thus allowing both sample designs to be compared under the same assumptions. To represent more accurately the probabilistic sampling in practice, two parameters were included in this model: K 1 and K 2, which determine how many attempts will be made by the interviewer to make contact with the selected household and with the selected respondent, respectively.This comparison will be done using the mean square error (MSE) and the time it takes to finish the collection of data (number of contacts). Different estimators of the probability of response for each of the studied sampling designs are compared. Time allowing, I will also present an empirical assessment of the quality of the prediction of 898 electoral surveys conducted in Brazil between the years 1989 and 2004.

An Empirical Evaluation of Respondent Driven Sampling from a Total Survey Error Perspective

Zeynep Suzer-Gurtekin[1], Sunghee Lee[1]
University of Michigan[1]

RDS is a sampling and data collection method that has been widely employed in data collection for rare and hidden populations. However, the practices of RDS and underlying assumptions required for statistical inferences are yet to be understood clearly by survey researchers who rely on traditional probability samples. This paper relies on the total survey error frame and focuses on the sampling process of RDS and how sampling error in RDS can be affected by other error sources, mainly nonresponse and measurement errors. We describe and present empirical evaluations of these errors in RDS. The study uses publicly available datasets: 1) the Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) conducted from November 2006 to August 2008, and 2) Latino MSM Community Involvement: HIV Protective Effects conducted in 2004.

The Utility of Weighting Methods for Reducing Errors in Opt-in Web Studies

Jill Dever[1], Bonnie Shook-Sa[1]
RTI International[1]

Probability sampling designs, those with samples selected with a reproducible random mechanism, are considered by many to be the gold standard for surveys. Theory has existed since the early 1930's to produce population estimates from these samples under the labels such as design-based, randomization-based, and model-assisted estimation. This theory ultimately requires that the sample units excluded from the analysis files either because of non-sampling or nonresponse, are missing at random. This situation, however, is not always attainable. || Studies involving samples without a necessarily reproducible sample design, referred to as non-probability surveys, have gained more attention in recent years but they are not new. Touted as cheaper, faster (even better) than probability designs, these surveys capture participants through various methods such as respondent-driven sampling or opt-in web surveys. For surveys required to produce population estimates to meet their stated fit for purpose, the link between the sample and the target population as well as the probability of participation must be addressed to justify the desired level of quality. Survey weights or analytic models can provide the needed evidence of the data's utility but the research suggests that the results are inconsistent. || In this article, we first review the weighting methods currently in use for web opt-in and other non-probability surveys. Next, we describe a simulation study to directly compare these methods and summarize the findings related to error reduction. A specific example of an opt-in web study on smoking cessation techniques is used to ground the discussion. We conclude with a few recommendations with an eye toward the study's fit for purpose context and our work on a quality framework for probability and non-probability surveys.

Probability Samples – Meet Your Match! A Comparison of Two Distance Measures for Linking Non-probability and Probability Based Samples

Trent Buskirk[1], David Dutwin[2]
Marketing Systems Group[1], SSRS[2]

Non-probability based opt-in samples and panels are beginning to emerge in popularity in many research areas including marketing, consumer, social and political sciences. Some research has explored methods for reducing the impact of self-selection bias on the overall estimates derived from non-probability samples, including simple post-stratification calibration, propensity score adjustments and sample matching. Yet there is almost no literature to date that considers the relative merits of these approaches from a Total Error perspective. Post-stratification adjustments rely only on the non-probability sample (and some external data source(s) providing benchmarks) but propensity score adjustments and

sample matching rely on both the non-probability sample as well as a probability sample. In particular, sample matching identifies a subset of the non-probability sample that is linked to units within a temporally relevant probability sample using a distance function measured on key indicators, or more generally bases the match on a propensity score. Analyses and estimates are then produced using the matched sample subset of the non-probability sample. Nearest neighbor distance functions have been commonly used as the basis of matching variables, but this method, in theory, can limit the number of candidate variables for the match and suffer from the "curse of dimensionality problem." In this study we explore the use of random forests as the basis of generating a matched sample using the proximity measure generated from applying the unsupervised version of random forests to the combined data set consisting of a probability sub-sample and the non-probability sample. Sample matches are generated by selecting the member from the non-probability panel that has the highest proximity measure for each of the members in the probability sample, where ties are broken randomly. We compare the resulting matched sample to one obtained using the more common nearest neighbor approach. The number of variables used for matching will also be varied to demonstrate possible advantages of the random forest models compared to nearest neighbors. Data for our study come from a random sample selected from an ABS sampling frame along with additional sample obtained from an online, opt-in non-probability panel. Key outcomes of interest involve media related measures and predictors for the sample matching include a battery of demographic variables. We evaluate final estimates obtained from the matched samples defined from each of the methods using an error framework that incorporates both bias as well as variance. Overall, the principal purpose of this research is to consider the fusion of probability and non-probability samples and compare two methods for generating matched samples within a Total Survey Error framework.

Teaching TSE & Big Data: Presentations and Roundtable Discussion

Session Chair: Beth-Ellen Pennell

The Survey Octopus: Creating Better Conversations About Total Survey Error with Non-specialists

Caroline Jarrett[1]
Effortmark Ltd[1]

Although the concepts of Total Survey Error (TSE) have been widely accepted amongst survey methodologists and statisticians for many years, TSE is still not widely understood by people who commission ad-hoc surveys for business or government. After having many conversations with colleagues and clients where I tried to explain that a high number of responses was in itself not a guarantee of data quality, I turned to the classic Survey Lifecycle (Groves et al, 2009). In this presentation I will show how I evolved the Survey Octopus, a way of helping non-specialists to get to grips with the issues involved in TSE and to help them to make better, and more informed, choices when deciding about how to approach a survey. The Survey Octopus may also be useful as a way of introducing students to TSE, particularly those who are taking a class in survey methodology as a one-off.

Survey Methodology Courses and TSE/Big Data Issues. Classroom Experiences Among University Instructors

Wojciech Jablonski[1]
University of Lodz[1]

In Europe Survey methods are usually taught during standard methodological courses organized for students of sociology. The participants are expected to learn how to carry out research project utilizing survey methods. Moreover, they are instructed how to evaluate the survey data, how to assess the quality of different survey results. Although the importance of methodological knowledge concerning survey methods is often emphasized, a thorough literature review shows that the range of publications of this relevant topic is rather limited. In general, only few studies examine university teachers' opinion of their work, specifically the difficulties they encounter while conducting classes with students, and the solutions they implement in order to overcome these problems (Paino et al. 2012; Chin 2002; Baker 1985). The presentation outlines the selected results of the study conducted among Polish university instructors teaching survey methods. This research, a web-based questionnaire (which is going to be completed in May-June 2015), aims to assess the extent to which the issues concerning the TSE paradigm and Big Data are incorporated to the survey methodology courses. For instance, it delivers answers to the following research questions: Are students of sociology familiarized with the accuracy and the value of online panels for completing surveys (Callegaro et al. 2014)? Do the students learn how to use paradata to monitor fieldwork activity and analyze different errors (nonresponse, measurement, coverage) in surveys (Kreuter et al. 2013)? How deep is students' knowledge on different modes of data collections used in surveys and on advantages and limitations of mixed mode surveys (Dillman, Smyth, and Christian 2014; 2009; Dillman 1978)? Moreover, the project is

focused on the pedagogical problems the teachers experience while conducting classes with students. What kind of teaching techniques do the respondents use in order to overcome the difficulties? Are the teachers the authors of these techniques, or did they familiarize with them while participating in a pedagogical course/reading a pedagogical book, etc.? What kind of solutions should be – according to the teachers – introduced so that teaching or studying quantitative methodology, especially the TSE paradigm/Big Data, is a more pleasant and effective experience?

Introducing an International System of Online Courses in Survey Methodology and Big Data
Frauke Kreuter[1]
University of Maryland[1]

Mobile Surveys

Session Chair: Edith de Leeuw

Invited Presentation: **Smartphone Participation in Web Surveys: Choosing Between the Potential for Coverage, Non-response and Measurement Error**
Gregg Peterson[1], *John LaFrance*[1], Jamie Griffin[2], *JiaoJiao Li*[3]
ISR[1], *University of Michigan*[2], *Market Strategies International*[3]

Web surveys are increasingly completed (or at least begun) on small-format mobile devices (e.g. smartphones), regardless of the intention of the researcher or the design of the instrument. Mario Callegaro was among the first authors to bring this to researchers' attention and many others have since documented this trend (Callegaro, 2010; Peterson 2012; Comer & Saunders 2012; Jue 2012; Kinesis 2012). The growth in smartphone survey taking tracks well with Americans' increasing use of cell phones to access the internet (Duggan, Smith 2012), which in turn, coincides with the growth in smartphone penetration in the US. Should researchers' allow or disallow smartphone survey taking on surveys originally designed for larger screens? They do have a choice. This poses an interesting dilemma from a total survey error perspective. The decision requires a trade-off between non-response and measurement error. We know from Pew tracking studies that more than 20% of cell phone owners go online mostly on their phones (Duggan, Smith, 2013), and we might infer that many of these people spend little or no time reading email or browsing the internet from large-format devices. In fact, Pew also reports that 10% of Americans only means of at home internet access is via their phones. (Zickuhr & Smith, 2013) So allowing smartphone participation in web studies should help improve response rates among the part of the population that primarily or only accesses the internet from smaller devices. This is important because they are disproportionately younger, less-educated, less-affluent and non-white according to Pew. However, allowing smartphone participation for all but the simplest of survey instruments has the strong potential to introduce both non-response (via survey break-offs) and measurement error. A post-hoc analysis of unintentional smartphone survey taking demonstrates that compared with survey taking on larger devices, survey lengths can increase by as much as 50%, break-off rates are nearly doubled for those who begin on a smartphone device, and mode effects and measurement error are possible (Peytchev and Hill, 2010; Bailey and Wells, 2012; Peterson 2012; Grenville 2012; Jue 2012). Rich literature supports the notion that survey interface design (independent of the questions asked) can impact survey results. Couper, Tourangeau, Conrad and other colleagues (2004, 2007 and 2008) have conducted a number of web experiments in which various elements of the web survey interface have been altered or manipulated in test and control situations. Inconsistent column widths, uneven spacing of response options, contrasting images, background color variations, and slider bars as alternatives to radio buttons (among many others) all have been demonstrated to impact survey responses. "Respondents make inferences about the meaning of survey items based on visual cues such as the spacing of response options, their order, or the grouping of questions. These inferences affect how quickly respondents answer the questions, which answers they select, or both." While plain HTML surveys designed for PCs usually render fine on smaller devices, the surveys are often difficult to read and answer without frequent zooming (to make response buttons larger or even to read text), pinching and extra scrolling. The increased survey lengths and break-offs probably relate to both the design and usability of the surveys and devices, along with the natural lag on most cell phone networks or the speed of the particular networks where respondents browse. When allowing or disallowing mobile phones in surveys, robust participation seems to suffer either way. Disallowing smartphones may reduce participation by younger, more urban, and less affluent respondents who may never complete a survey on PC. On the other hand, allowing smartphones could lose these same people at higher than expected rates due to break-offs. Perhaps most concerning—independent of who chooses to participate or complete surveys—is that the devices themselves (and how our surveys render on them) may impact how respondents answer survey questions and introduce measurement error. In this paper, we attempt to demonstrate the non-response impact of disallowing smartphones, the futility of relying on warnings and recommendations about device usage, and we demonstrate how a well-chosen "mobile friendly" interface design can minimize the impact of non-response and measurement errors in web surveys that allow smartphone participation. We propose that in order to minimize Total Survey Error, researchers should allow smartphone participation on most surveys but simultaneously work hard to minimize measurement and non-response error by focusing on survey content, question types and interface designs. Specifically, we: 1) Review the literature from peer reviewed journals (as well as from other papers and conference

presentations) on unintentional smartphone survey taking as well as the literature on cell phone internet access in general. We review the evidence that unintentional smartphone participants (like cell mostly/only internet users in general) are demographically unique and that their survey experience is different with regard to survey length and breakoffs. We use this data to hypothesize about the potential for non-response error and bias, although we cannot measure it directly. Given the unique characteristics of smartphone participants, a survey design which actively disallows smartphone participation will clearly force non-response among respondents who only access the internet via the phones. Note: One could argue that this might also be categorized as a coverage problem as some portion of the internet population only has access via a cellphone. In other words, it could be thought of in the same light as the coverage error which occurs when we exclude cell-only households in telephone studies (e.g. Boyle, Lewis & Tefft, 2009, Peytchev et. al. 2010). Researchers who choose to include a cell-phone frame in household telephone studies, do so to not only cover the cell-phone only population, but also to improve response among the cell phone mainly population whose lack of participation in traditional landline only samples can contribute to effective coverage error. (Boyle, Lewis, and Tefft 2009). However, since most internet study sampling frames do not (or cannot) explicitly disallow smartphones; we will consider this as researcher induced non-response. 2) We introduce the results of a large experiment to discourage smartphone survey taking on a short customer service feedback survey. We find that discouraging smartphone participation and encouraging PC or tablet participation has little impact on the device choices respondent make. This inability to persuade smartphone participants to switch devices provides further evidence that a PC only survey may miss some important elements of the population and introduce non-response error. This is only directional evidence. Our experiment does not directly measure non-response bias in this experiment. 3) We compare self-selected smartphone participants versus PC participants on key response distributions in a study of university students. The otherwise demographically similar students who completed the survey via their smartphone device responded differently than their PC/tablet counterparts on measures important to the study. It is impossible to disentangle differences that result from the chosen mode of completion (i.e. a smartphone or a PC) versus differences arising from the types of people who chose to participate via a smartphone, however, these results suggest that disallowing smartphone survey taking has the potential to lead to bias. In this study, the differences were significant but not great enough relative to the size of the smartphone device user sample (~10% of the total) to change the results, even after completely removing smartphone participants from the analysis. However, the upward trend in smartphone internet access would suggest the issue will only grow in importance over time, much like the cell-phone only phenomenon has grown in importance for telephone researchers. 4) We present the results of a randomized experiment designed to compare response distributions for smartphone and non-smartphone survey participants. Our experiment also compares multiple "mobile friendly" design alternatives to learn which designs improve completion rates, survey length, self-reported user experience and respondent engagement while minimizing measurement error. By "mobile friendly," we mean designs that use larger fonts and larger interface elements, and still size to the available screen real estate on a smartphone. We randomly assigned respondents to complete the same survey using one of eight survey presentation alternatives—six smartphone alternatives and two PC alternatives. We found that all "mobile friendly" designs improve completion rates, survey length and user experience. We found very few differences in measures of engagement such as straight lining, speeding and satisficing in any of the alternatives. Differences on means and proportions on key ratings questions in this study represent the measurement error which can be introduced when allowing smartphone participation. We compare the results of 5 unique "mobile friendly" treatments versus a traditional PC treatment and show that only one of the designs consistently matched the response distributions of our traditional PC treatment. Interestingly, respondents who were assigned to a non-optimized smartphone interface (i.e. not "mobile friendly") also closely matched the PC treatment on key measures; however they broke off at significantly higher rates and had lower levels of self-reported satisfaction with the survey experience. Finally, we see evidence of potential non-response error by observing differences on selected frame variables between those who competed surveys on a smartphone device and those who broke off before completion. Given that smartphone survey takers breakoff at much higher rates than PC/tablet survey takers, we infer that the choice to allow smartphone may come at a price of increased non-response error in addition to measurement error.

Invited Presentation: Text Interviews on Mobile Devices

Frederick Conrad[1], Michael F. Schober[1], Christopher Antoun[1], Andrew L. Hupp[1]
ISR[1]

As people's communicative habits change with the rise of mobile devices, collecting survey data via SMS text messaging is an increasingly attractive alternative to traditional telephone interviewing. But the properties of text interviews are not well understood. Our primary focus in the proposed chapter is the measurement error properties of text interviews. We will also examine (and mostly reject) the possibility that the data quality advantages we observe for text versus voice interviews can be explained by patterns of nonresponse and breakoffs. In addition, the chapter will explore mobile and smart phone ownership (required for texting) by examining recent Pew data in order to explore whether coverage error might affect the accuracy of estimates based on text surveys. The clearest coverage gaps exist for older members of the public, and in the case of smart phones, for those with the least education. Most of the chapter will be devoted to reporting a study we conducted in which 634 iPhone users answered 32 questions in one of four modes to which they were randomly

assigned: text interviews administered by either human interviewers or an automated text-interviewing system, and voice interviews administered by either human interviewers or a speech IVR system. The main finding is that responses are of higher quality when collected in text than voice interviews. Text respondents provided more precise numerical answers (fewer rounded answers), straightlined less, and disclosed more sensitive information than respondents in voice interviews. We attribute the greater precision of text responses to the reduced time pressure created by its asynchronous character – respondents don't feel they need to respond at the rapid fire pace that spoken interviews seem to demand. Being able to respond when they are ready to allows text respondents to answer thoughtfully, and possibly to consult records. We attribute the increased disclosure observed in text interviews to reduced evidence that an interviewing agent (human or automated) is present: no voice, no paralinguistic behavior, no judgment. Finally, at least as many text respondents were very or somewhat satisfied with their interviews as voice respondents, even though text interviews lasted far longer. In the chapter, we will also consider the possibility that different rates of nonresponse and breakoffs might account for data quality levels in text interviews: more conscientious sample members (i.e., those less likely to round or straightline) or those more willing to disclose sensitive information may be less likely to start or finish voice than text interviews, which could have produced our observed patterns. In our sample of volunteers, all of whom had agreed to participate in an iPhone survey before they were randomly assigned to an interviewing mode, this explanation is not supported: we see no evidence that age, race, education, income or ethnicity differed between respondents and nonrespondents nor between completers and non-completers. Although in our data set nonresponse and breakoff explanations are not supported, text interviews may well reduce the likelihood of noncontact: the visual persistence of an invitation in the iPhone's messages app suggests that a nonrespondent has read and rejected it. Perhaps this is why we observed higher response rates in text than voice interviews – one of several reasons why text is a promising survey mode.

Invited Presentation: **Mobile Web Surveys: A Total Survey Error Perspective**

Mick P. Couper[1], Christopher Antoun[1], Aigul Mavletova[1]
ISR[1]

Surveys completed on mobile devices (especially smartphones) are increasingly common, whether intended by the survey designer or not. The growing literature on this topic suggests serious challenges for designers of mobile Web surveys, whether app-based or browser-based, whether optimized for smartphones or not. For example, evidence suggests that those using mobile devices to access Web surveys have significantly lower response rates, higher breakoff rates, and longer completion times than those using a PC. We propose a chapter that offers a framework for understanding the impact of the mobile revolution on Web surveys from a total survey error perspective. We would examine the implications on the mobile revolution for self-administered data collection (i.e., we are not focusing on cell phone and telephone surveys). For instance, mobile Web raises the prospect of using RDD methods to sample mobile phone users and invite them using SMS (or text messaging), affecting sampling error. Similarly, the penetration (and use) of the Web on mobile devices is still uneven across countries and across demographic groups, with implications for coverage error. A big challenge for mobile Web surveys is nonresponse error. Most of the work has focused on observed response rate differences between those who use mobile devices versus PCs to access Web surveys, but our research is focused on understanding the reasons behind the differential nonresponse (i.e., nonresponse bias) and ways to mitigate these effects. Mobile apps offer a good example of the trade-off between nonresponse error and measurement error. Requiring sample persons to download and install apps may increase coverage and nonresponse error but may improve the quality of measurement and the survey experience for those who do so. Similar trade-offs exist with regard to the use of mobile devices for unobtrusive data collection (e.g., GPS tracking) and frequent measurement (e.g., ecological momentary assessment). Much research (including our own) is focused on optimal design for mobile Web surveys to minimize measurement error. Optimizing a survey for mobile Web may reduce some of the problems experienced by mobile users, but may introduce differences between the two designs. In summary, this proposed chapter would offer a theoretical framework and review the small but rapidly growing literature on mobile Web surveys, focusing on various sources of survey error. Collectively, we are making substantial contributions to the mobile Web literature. For instance, Antoun and Couper (2013) are examining coverage issues of surveys that include only mobile users or include only PC users. Antoun is currently collecting data on the LISS mobile Web panel to explore both nonresponse (see Antoun, 2014) and measurement error issues. Mavletova (2013) has conducted experiments on data quality in mobile Web surveys, and Mavletova and Couper (2013a, 2013b, 2014) are collaborating on research on both nonresponse and measurement error issues. In addition, we are familiar with the research in this area, and in touch with others working on mobile Web surveys. We are well-positioned to address this topic. This chapter would thus draw heavily on our own research, while also synthesizing the research of others on the topic.

Big Data from TSE Perspective

Session Chair: Peter Lugtig

Invited Presentation: Decomposing Twitter from a Total Survey Error Perspective

Joe Murphy[1]
RTI[1]

The past decade has witnessed the rise of social media as a popular means of communication. This rise has coincided with a continued general decline in the public's willingness to take part in survey research. With constrained research budgets and an increasing importance put on the timeliness of results, many researchers have begun investigating social media as a potential source of information from which to capture some of the same opinions and behaviors traditionally addressed through surveys. One platform in particular has received considerable attention as a means of passively analyzing trends in health, politics, and other topics—Twitter. A social microblogging platform, Twitter allows users to post updates ("Tweets") of up to 140 characters to let others know "what's happening." Over half a billion Tweets are posted each day and these data are available to researchers looking to capture sentiment or discussion without any interaction with the poster. Several studies in recent years have claimed to replicate survey findings using this alternative source, suggesting, perhaps, a breakthrough in the ability to supplant the designed data of a survey with organic big data. For example, Broniatowski et al. (2013) report 85% accuracy in replicating the Centers for Disease Control's survey-based weekly change in direction of influenza prevalence in the United States using the automated processing of Tweets. The authors claim their system can provide similar results in a fraction of the time and cost of the survey of outpatient providers. With the enthusiasm surrounding the potential of this new data source, some researchers have essentially forged ahead without a careful consideration of the potential error sources that should be investigated when conducting passive analysis using Twitter. As the central organizing structure of the field of survey methodology, the Total Survey Error framework can provide a starting point and allow researchers reared in survey methods to better understand and decompose Twitter as a research resource. For example, Twitter is subject to a variety of both observation and non-observation error. It can be viewed, in some ways, as a large opt-in "survey" with no control (or influence) of the researcher with regard to question wording, reporting format, or the number of responses from a single individual. Of course, with such a different purpose and structure, it is likely that Total Survey Error will not sufficiently capture all areas of concern with Twitter, nor will all components of Total Survey Error be applicable to Twitter. However, beginning with the basic components of specification, frame, nonresponse, measurement, and processing, we can begin to evaluate Twitter and surveys under a common set of considerations. This chapter will present several examples in which Twitter has been considered as an alternative to a survey and will objectively decompose the error sources to lay the groundwork for future conversations of the quality and costs of each. From this starting point, we can consider a mutually exclusive and exhaustive framework and set of metrics that could possibly be applied to other types of social media with a few modifications.

Invited Presentation: Big Data: A Survey Research Perspective

Reginald Baker[1]
Private Consulting[1]

Big data is one of those terms that can mean different things to different people. To some it simply means datasets so large that massive computing power is required to process them. To others it refers to the exponential growth and availability of all kinds of data. In this more intriguing view big data means the merging of data from three principal sources: (1) customer data, that is, the tracks we leave behind each time we buy something; (2) the mostly unstructured data of social media; and (3) the Internet of things, meaning the increasing use of smart objects—mobile phones, appliances, cars, etc.—capable of measuring and transmitting information about how and where they are being used. Researchers, especially market researchers, are enthusiastic about big data because of the potential it offers not only to study behavior but also to build models to predict it. Some even see it as a potential substitute for surveys, which have become increasingly difficult, expensive, and sometimes less reliable. One common cliché is that big data can tells us "what" and surveys tell us "why." With the rise of behavioral economics and a greater appreciation of the role of emotions and intuition as drivers of behavior (system 1, system 2, and all that) it is no longer clear that we even know or can express why we did x or why we did y. So asking questions may not be the most effective way to understand behavior. And as text processing algorithms mature, there may be still another alternative to survey research, one that relies on the unstructured data of social media. At the same time, big data enthusiasts generally recognize the challenges it presents, often expressed as "the 3Vs" model: volume (amount of data), velocity (speed of generation), and variety (range of data types and sources). Market researchers have begun to talk about a fourth V-veracity. Simply put, does this amalgam of data in front of me represent what it purports to represent? Not surprisingly, they do not think of this in terms of TSE, but the questions that need to be asked are in many ways the same. This paper will explore big data from a TSE perspective. It also will consider how

lessons learned from our increased use of non-probability sampling techniques may be preparing us to understand and use big data more effectively. Finally, it will speculate about the future of big data and what it may mean for surveys and the survey profession. Key here is an understanding of the data science paradigm, how it differs from the ways in which most of us now approach the analysis of data, and whether the two perspectives can be reconciled.

Creating a New Variable as a Means of Assessing the Item External Validity by Using Big Data

Andrei Veikher[1]
NRU HSE - St.Petersburg[1]

Big Data techniques allow to obtain new comprehensive data on a larger circle of events and facts of the target groups life. This creates favorable conditions for the TSE methods to be supplemented by external estimates of local errors. Traditionally, external validity criteria were simply socio-demographic indicators: gender, age, ethnicity etc. Distributions on these indicators were compared with data from independent sources. The discrepancy between them was the basis for the use of weighting methods in order to refine distribution of all the other variables. Today we have a level of non-response often exceeds 50%. The reasons for this phenomenon are include a range of social and cultural factors. Their relationship with the main socio-demographic indicators is ambiguous. Each of them can distort the sample on the part of the studied indicators, without affecting other. This paper offers a look at Big Data independent indicators related to individual topics surveys that have several indicators in the questionnaire. Such indicators are translated into a new variable, for which there is an adequate indicator of an independent source The difference between the distributions of the new variable and an external indicator is the criterion validity of the survey on a particular topic. In our surveys such external indicators were nominal wages in the region, the number of retirees who have privileges to public transport, the number of residents who visited the clinic last month. The questionnaire provides two-four questions on the basis of which it was possible to calculate a new variable, similar to the named indicator.The statistical reliability of this estimate depends on Bayesian. The method "item external validity of sample" was used in studies of shadow wages, unreported trips on public transport, latent disease in Saint-Petersburg (2001-2011). Creation of new registers storing indicators of population gives hope to expand the range of applications of this approach.

Examining Big Data in the Total Survey Error Framework: A Synthesis of the Current Research

Celeste Stone[1], *Cong Ye*[1], *Ahmad Emad*[1]
American Institutes for Research (AIR)[1]

While big data has been around for years, the recent increases in its availability and accessibility for research studies has shifted big data into the spotlight. Big data seems to offer unlimited possibilities to many survey practitioners, however, even those in the commercial sector have found that using big data for research requires a large initial investment to gather, process, and analyze big data. In a field committed to meticulously examining data in terms of total survey error (TSE), survey methodologists are struggling to find ways to use big data to either supplement or replace survey data because so little is still known about the big data's characteristics with respect to sampling and nonsampling errors. At the 2014 International Total Survey Error Workshop, Paul Biemer offered some suggestions for applying and adapting the TSE paradigm for big data. However, the field has just begun empirically evaluating big data—or more accurately "found data"—in the total survey error framework. As reported in the forthcoming [2015] AAPOR task force report on big data, it is not the size of the data, but rather the "found" nature of the data that is of primary concern to survey researchers. What has been learned so far about found data's characteristics in terms of total survey error, as well as its fitness for use? In what aspects and to what extent have survey researchers used big data to improve coverage, sampling error, nonresponse, and measurement error in the empirical research? How (what techniques) are researchers using big data to improve "traditional" survey data? This paper reports the findings of a systematic review of research that empirically evaluates found data in the total survey error framework. It will provide a necessary first step in helping survey researchers decide if and how found data can be used to supplement or replace survey data. We will not only synthesize the results from existing studies in this area but discuss the implications these findings have on these data's use in survey research. In the process, we will suggest in what direction future research should go and identify possible datasets that could be used in this endeavor.

Estimating and Adjusting Survey Errors in Mixed-Mode Data, Part II

Session Chair: Thomas Klausch

Estimating Components of Mean-Squared Error as a Means to Evaluate Mixed Mode Solutions to Noncoverage Error in Telephone Surveys

Caroline Vandenplas[1], Caroline Roberts[2], Rosa Sanchez Tome[3], Michèle Ernst Staehli[4], Dominique Joye[3]
Center for Sociological Research, University of Leuven, Leuven, Belgium[1], University of Lausanne, Switzerland[2], University of Lausanne[3], FORS, Switzerland[4]

In Europe, mixed mode data collection is increasingly being adopted by survey organisations looking to address growing noncoverage errors associated with telephone surveys. Despite the growing popularity of mixed mode surveys, however, there are still relatively few studies evaluating their claimed advantages over traditional, single-mode studies. Yet there are good theoretical and practical reasons for concern about the enthusiasm with which mixed mode surveys has been met. Perhaps the most important of these is the potential increase in Total Survey Error associated with using a supplementary mode (or modes) to try to offset the limitations (e.g. noncoverage error, or nonresponse error) of a primary mode. Each additional mode introduces a confounding of measurement and selection errors, rendering the data from different parts of the sample non-comparable. The implications of this for data complexity mean that mixed mode data impose considerable burden on analysts, who may be unaware or unequipped to deal with the problems introduced by the survey design. These drawbacks of mixing modes, in addition to the considerable costs involved in mounting survey fieldwork in multiple modes (despite the promise of so-called 'sequential' designs, which encourage sample members to respond in more cost efficient modes, while reserving more expensive contact and interview strategies for nonresponse follow-ups), highlight the need to use suitable metrics for assessing the total quality of different types of survey design, so that mixed mode alternatives can be appropriately compared with traditional data collection protocols.

Mode Effects in American Trends Panel: A Closer Look at the Person-Level and Item-Level Characteristics

Stanislav Kolenikov[1], Kyley McGeeney[2], Scott Keeter[2], Courtney Kennedy[1]
Abt SRBI[1], Pew Research Center[2]

This presentation is based on the American Trends Panel by Pew Research Center and Abt SRBI. ATP is a probability panel with RDD recruitment. The panel currently has 4,165 recruited active members, of whom approximately 3,200 complete a typical wave. Panel surveys have been conducted on different modes in different waves, including web for most panel members, and mail or phone for those who do not have access to the Internet. We analyze the results of the July wave (Wave 5) that included a comprehensive, large-scale mode-of-interview experiment that randomly assigned respondents to telephone and web modes, with approximately 1,500 respondents in each mode. A 75-question instrument that included of a variety of question types and topics was administered. For the purposes of mode effect analysis, each experimental group was weighted separately to national parameters for the general public. To quantify the contributions to the mode effects of the different question characteristics, we build a regression model with effects of person and question characteristics to identify the properties of survey questions that make them susceptible to mode effects, as well as the demographic groups that tend to exhibit mode effects. The question characteristics are coded using a scheme enumerating several properties of questions such as type of question (attitude, behavior, knowledge) or its format (yes/no, unipolar, bipolar, frequency, etc.). We discuss how the decomposition of the total survey error and explained variance helps identifying the properties of the questions that are associated with the mode effects, such as question format, topic, and the potential impact of social desirability.

Comparison of the Quality Estimates in a Mixed-mode and a Unimode Design: An Experiment from the European Social Survey

Melanie Revilla[1]
RECSM, Universitat Pompeu Fabra[1]

In the frame of the European Social Survey (ESS), a series of experiments were conducted to investigate if and how the ESS might move from the single face-to-face survey to a mixed-mode design. In order to determine this, many aspects have to be considered. As the ESS wants to maintain the possibility to compare its data across countries and across time, one of the requirements to introduce a mixed-mode design is that it leads to a similar data quality as the current unimode face-to-face design. In this study, we define the quality as the strength of the relationship between the latent concept of interest and the observed answers. Analyzing the experiment done in parallel of the ESS round 6 (2012–2013) in Estonia and the UK, we find that the quality is similar in the unimode and mixed-mode designs, at least for given scales. This is true both for single items and for composite scores. Therefore, standardized relationships in the main ESS round 6 and the mixed-mode experiments can be compared. Besides, for the composite scores, we also find metric and scalar invariance, meaning that unstandardized relationships and means can be compared for the two concepts tested across the unimode and the mixed-mode designs too.

Mode Preference as a Covariate for Estimating Mode Effects

Caroline Vandenplas[1], Jorre Vannieuwenhuyze[2]
Center for Sociological Research, University of leuven, Leuven, Belgium[1], University of Utrecht[2]

Mixed-mode surveys, in which different respondents complete the survey by different survey modes, become increasingly popular in an attempt to lower survey cost, respondents burden and coverage problems. Each mode induces possible confounded selection and measurement effects that can render the interpretation of the estimated parameters as well as the comparability with other single or mixed mode survey delicate. Measurement and selection effect can be estimated by calibrating the mode groups on a set of covariates, which are often chosen to be socio-demographic variables within the existing literature. However, the covariates must meet two important assumptions. They must be mode-insensitive and fully explain selection effects between the modes. Especially the second assumption might be problematic for socio-demographics and require the investigation of alternative covariates. One example of alternative covariates is questions about mode preferences. Such questions may provide a better trade-off between both assumptions because they may better explain why different respondents answer by different modes. Calibration by mode preference and socio-demographic variables are discussed in this paper and illustrated by Estonian European Social Survey data, which yields slightly better overall results for calibration by mode preference compared to socio-demographic covariates. The results point at a need to explore alternative calibrating covariates for analyzing mixed-mode data.

Adaptive Design

Session Chair: Meena Khare

Invited Presentation: **The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth: Results from a Randomized Experiment**

Brady West[1], James Wagner[1], Heidi Guyer[1], Frost Hubbard[1], Jennifer Kelley[1], Mick Couper[1], William Mosher[1]
ISR[1]

The National Survey of Family Growth (NSFG) is a national cross-sectional study of females and males aged 15-44. Data are collected utilizing a continuous data collection model in which an average of 5,000 interviews is completed annually over the course of four quarters. Like many other national U.S. surveys, response rates in the current NSFG cycle (2011-2019) have been declining slightly relative to the most recent cycle (2006-2010). Empirical evidence suggests that the declining response rates have largely been a function of 1) decreasing cooperation rates among identified eligible persons, 2) higher average counts of calls per completed case, and 3) higher proportions of eligible individuals who did not express serious concerns but still chose not to participate. Interview length has also been slowly increasing over this time. In response to these trends, we implemented a randomized experiment aimed at appealing to the extrinsic motivation of sampled persons by increasing the NSFG "token of appreciation" from $40 to $60. Beginning with the ninth quarter (late 2013), area segments were randomly assigned into one of two groups: the $40 incentive or the $60 incentive. Advance mailing materials differed minimally for the two groups, and the interviewers working the area segments in a given primary sampling unit were responsible for communicating the incentive amount. This study will use the Total Survey Error (TSE) framework to examine the effects of this randomized experiment on a variety of cost and error indicators for a one-year period from September 2013 through August 2014. We will first consider sampling error, comparing overall sample yield between the two groups in both the first and second phases of data collection. The second phase focuses effort on a subsample of active nonrespondents after 10 weeks in a given quarter, and increases the incentive to $80 in each group. Next, considering nonresponse error, we will compare contact rates, eligibility rates, cooperation rates among eligible persons, rates of resistance, and "screen-and-go" rates (instant interviews after initial screening interviews) between the two groups. Because a primary goal of the experiment is to increase participation among specific subgroups of respondents, we will also compare the two groups in terms of 19 key statistics, and determine whether any differences in response rates are also accompanied by differences in distributions for these key variables. Third, considering measurement error, we will compare the reliability of sensitive survey measures collected using both CAPI and ACASI. Finally, we will compare the average number of contacts and calls per completed case, to determine whether the increase in costs engendered by the higher incentives may be offset by decreases in the amount of effort necessary to obtain interviews. All analyses will also be performed for key population subgroups defined by gender, race/ethnicity, and age. Overall, the results from this study will help managers of national U.S. surveys to determine whether an increased incentive might alter current declining trends in response rates, and whether the changes may also have positive or negative impacts on other important cost and error indicators.

Targeted Sampling, Mixed Mode, Incentives, and Paying for Completion: What Works for Reaching Hard to Survey Low Income Households with Civil Legal Needs?

Danna Moore[1]
Washington State University[1]

Low income families are hard to reach and it takes extra effort to identify, and entice these households to surveys. The goal of the research was to test Address Based sample frame attributes and mixed mode survey methods for reaching and identifying households that met low income thresholds and determine the extent of their civil legal needs. This study is important to understanding the circumstances of low income households and expand public support for civil legal aid. Mixed mode survey methods (telephone, mail, and web) and experimental testing of differing amounts of token cash incentives, high and low valued lotteries, and payment of $20 for completion are used to draw respondents to the survey. This study is significant as it tests new methods of using ABS sample targeting low income census tracts and survey techniques, using a combination of upfront cash incentive with promise of payment for survey completion, for effectively garnering participation for low income individuals. It also tests the effectiveness of a novel income eligibility question. Pilot study results show that methods and survey modes make a difference. Cash incentives and lotteries were an important inducement to determine household income eligibility, projecting sample size for the full study and to obtaining sufficient surveys in the full study to analyze civil legal problems for households in WA State. This study forecasts the sample size and effort to get sufficient responses to support policy decisions.

Reducing Bias and Sampling Error: Using Simulation to Identify Effective Adaptive Design Strategies for the Crops Agricultural Production Survey

Herschel Lisette Sanders[1], *James Wagner*[2], *Jaki McCarthy*[3], *Ji Qi*[2], *Frauke Kreuter*[1]
RTI International[1], *University of Michigan*[2], *USDA-NASS*[3]

Nonresponse rates have been growing over time leading to concerns about survey data quality. Adaptive designs seek to allocate scarce resources by tailoring the survey design protocol for specific subsets of sampled units. These designs can be used to control multiple error sources. Adaptive designs often seek to identify subsets with two key features: 1) those whose probability of response can be impacted by changing design features, and 2) those cases who, once they have responded, can have an impact on estimates after adjustment. Subsets of cases that meet the first condition may help reduce sampling error for fixed budgets. Subsets of cases that meet the first and second condition may help reduce nonresponse error. The National Agricultural Statistics Service (NASS) is investigating the use of adaptive design techniques in the Crops Agricultural Production Survey (Crops APS). The Crops APS is a survey of establishments which vary in size and, hence, in their potential impact on estimates. In order to identify subgroups that have a large impact on adjusted estimates or their variances, we implemented a simulation that used Census of Agriculture (COA) data as proxies for similar survey items. Then, we simulated different patterns of nonresponse to identify subgroups who may reduce estimated sampling variance or nonresponse bias when their response propensities are changed.