CSDI Comparative Survey Design & Implementation

2014 International Workshop on Comparative Survey Design and Implementation
(CSDI 2014 – Bethesda, Maryland, USA)
Session Abstracts

Session 1 – Cognitve Interviewing

1. **Empirical Evaluation of Cognitive Interview Techniques**
   Patricia Goerman and Ryan King, U.S. Census Bureau

   In recent years, increasing amounts of empirical research has been done to examine the effectiveness of cognitive interview techniques (Presser, et al., 2004, Pan, 2008). However, little work has looked systematically at the functioning of specific cognitive interview probe wording, particularly in the context of translation to non-English languages. Typical cognitive interview probes include meaning oriented probes, process oriented probes, paraphrasing probes and think aloud procedures. This paper discusses a study of respondent reaction to traditional and modified Spanish-language cognitive interview probe wording. The study examines which wording appears to have elicited more useful respondent reaction to survey questions. Useful respondent reaction is defined as that which assists in the evaluation of whether survey items are being understood as intended. Cognitive interviews were conducted with 50 Spanish speakers with either lower than a high school level education (3/4 of respondents) or greater than high school (1/4 of respondents). The cognitive interviews tested a segment of the Census Bureau American Community Survey in the Computer Assisted Personal Interview (CAPI) mode. Two different types of interviews were conducted; structured interviews, which mimicked traditional English-language cognitive interview protocols translated directly into non-English languages, and experimental interviews, which were more open-ended with increased explanation upfront and flexibility of probe wording. Interviews were summarized and interviewer and respondent interactions were systematically coded using a modified behavior coding schema. This paper focuses on the creation of this coding system, which was used to evaluate cognitive interview probe functioning. It also discusses the application of the kappa statistic to evaluate inter-coder reliability in the context of the new coding scheme. The talk will conclude with a discussion of our preliminary recommendations for Spanish cognitive interview probe wording and procedures.

2. **A Comparative Study of English and Chinese Cognitive Interviews**
   Yuling Pan, Virginia Wake Yelei, and Grace Chan, U.S. Census Bureau; Gordon Willis, National Cancer Institute, National Institutes of Health

   Cognitive interviewing is a pretesting method frequently used to assess survey questions (Willis 2005). As the United States has become increasingly linguistically diverse, public and private survey organizations have begun to address the challenges of collecting accurate, quality data from residents who speak little or no English. As a result, in recent years, survey organizations have translated data collection instruments and other materials from English into multiple languages, and have adopted the cognitive interviewing method to pretest translated surveys in non-English languages. The usual practice for conducting non-English language cognitive interviews is to adopt the English cognitive interview techniques without much modification and

then translate the interview probes into target languages. This practice has confronted some challenges. Previous studies (e.g., Pan 2004, Pan et al. 2010, Carrasco, 2003, Goerman 2006) have documented some of the difficulties in conducting cognitive interviews in non-English languages. In spite of these efforts, no research thus far has focused exclusively on systematically examining how cognitive interview techniques perform across language groups and how effective they are in generating data for cross-cultural study.

The current study aims to assess the effectiveness of cognitive interviewing techniques in English and Chinese through a procedure that: (a) systematically categorizes subjects behaviors such that these can be quantified and comparatively analyzed; (b) provides a disentanglement of potential confounds (e.g., age, gender, acculturation level, language). In order to do so, we designed the study to include subjects from three groups: monolingual English speakers (n=15), bilingual English/Chinese speakers (n=30), and monolingual Chinese speakers (n=15). Half (15) of the bilingual subjects were interviewed in English, and the other half were interviewed in Chinese. Each group was also stratified according to gender, age, educational backgrounds, and acculturation levels. A total of 60 cognitive interviews were conducted with 30 in each language. Then we used a linguistic coding scheme (Pan 2013) to code the English and Chinese responses to interview probes to examine the effectiveness of these techniques. Results show that: 1) translated Chinese probes are not as effective as their counterparts in English; 2) linguistically, the forms of the probing questions in the source language do not necessarily match the functions in the target language; 3) On the cultural dimension, the probes elicit a different frame of interpretations. As a result, some of the probes don't elicit what they are intended to gather; and 4) the whole structure of interview (including the introduction and the background of the survey interview experience, etc.) plays an important role in the result. Findings from this research can be used by survey researchers to explore alternative c.

3. **Integrating Validity Evidence Based on Response Processes and Psychometrics in Cross-Lingual and Cultural Survey Research**
Jose-Luis Padilla, University of Granada, Spain

Validity evidence based on response processes was first introduced explicitly as a source of validity evidence in the latest edition of the Standards for Educational and Psychological Testing (AERA et al., 1999). On the other hand, equivalence has been widely considered the conceptual landmark to judge the validity of cross-lingual and cultural comparisons. However, there are no clear theoretical arguments and practical indications on how validity evidence of response process to survey questions and test items can contribute to the equivalence of cross-lingual and cultural comparisons. The aim of the paper is (1) to determine when it is critical to have evidence based on the response process to support the equivalence of cross-lingual or cultural comparisons; and (2) present a mixed-method framework for conducting validation studies that allow researchers to integrate validity evidence on response processes obtained by cognitive interviewing, and psychometrics in cross-lingual or cultural research.Together with a brief systematic literature review, theoretical and practical argument will be discussed. In addition, examples of validation studies conducted to combine validity evidence based on response processes with psychometrics will be summarized. Arguments for determining when validity evidence based on response processes is critical for supporting the use of the test or questionnaire in cross-lingual and cultural setting, along with indications on how to conduct a validation study aimed at integrating validity evidence based on response processes and psychometric will be discussed.

4. **Ways to Test for Equivalence: A Mode Comparison of Probing in Cognitive Interviewing and Online-Probing**
   Katharina Meitinger, GESIS

   Due to the growing significance of international studies, the need for tools to detect flaws in the comparability of questionnaires is pressing. Cognitive interviewing supplements statistical procedures in revealing problems of non-equivalence. Although this method helps with the analysis of items and the questionnaire design of future studies, its application in a cross-national context is challenging (Thrasher et al. 2011). Recently, the method of web probing has been developed that incorporates probing techniques in cross-national web surveys. It allows for a comparatively inexpensive increase in sample size and a quantification of results (Behr et al. 2012). However, differences between both methods might be expected. On the one hand, cognitive interviewing allows for a higher level of spontaneity and motivation. On the other hand, the large sample size of online-probing enables judging the prevalence of an interpretation pattern (Behr et al. 2013). The presentation fills a research gap by comparing probing in cognitive interviewing and online-probing. The results are based on items from the ISSP module on National Identity. Responses from 20 cognitive interviews and an online survey with 532 German respondents allow for a mode comparison. Data has been collected in April and September 2013. Both samples had a quota by gender, age and education and respondents to the online survey were drawn from a nonprobability online panel.

   The presentation compares both methods from different perspectives. First, the presentation adopts an error perspective (based on DeMaio & Landreth 2004) and shows which types of error online-probing and cognitive interviewing detect. Second, the presentation follows the research of Oudejans and Christian (2010) and assesses both methods in regard of the number of words/themes and the prevalence of additional explanations/examples. Third, the presentation also analyzes the role of spontaneity and the extent of its error identification potential. An assessment of the possibilities and limitations of both methods to improve international surveys concludes the presentation.

<u>Session 2 – Measurement Error</u>

5. **Assessing the Robustness of Established Construct Associations Across Race/Ethnic Groups**
   Tim Johnson, Allyson Holbrook, Marina Stavrakantonaki David Sterrett, Noel Chavez, and Saul Weiner, University of Illinois at Chicago; Sharon Shavitt, University of Illinois at Urbana-Champaign; Young Ik Cho, University of Wisconsin-Milwaukee

   An often overlooked consequence of survey measurement disparities across respondents from various race/ethnic groups is the selective attenuation of well-documented, established relationships between various sets of constructs. Of particular concern is the possibility that poor measurement of key constructs may lead to weakened, or absent, relationships between variables that prior research has documented. Alternatively, cross-group differences in construct associations may also represent culturally-driven differences in the social processes being examined. To investigate these alternatives, we compare a set of generally recognized construct associations across samples from four race/ethnic groups (African Americans, Korean Americans, Mexican Americans, and non-Hispanic whites) and across three interview languages. In addition to empirical findings, we consider the utility of this approach as an additional methodology for evaluating cross-cultural measurement comparability.

6. **Increasing Cross-national Comparability of Disability Measures with Anchoring Vignettes**
   Mengyao Hu, University of Michigan

   While assessing well-being is an important topic for cross-national studies, uncertain comparability of its measurements is an impediment. One source of measurement incomparability is the difference in response scale usage. Whether due to cultural influence in a cross-cultural research setting or due to instrument translation that produces less-than-perfect equivalence in a multilingual research setting, differential response scale usage across countries confounds the true differences in well-being in an unknowable direction. This study attempts to understand well-being across 11 countries through disability measured in five domains (pain, mobility, cognition, breathing, and affect) collected in multiple data sources: the Health Retirement Survey, the Survey of Health, Ageing and Retirement in Europe, the English Longitudinal Study of Ageing, and the Chinese Health and Retirement Longitudinal Study. Respondents self-rated their disability on the scale of none, mild, moderate, severe, and extreme. This scale is comprised of vague quantifiers, and its usage is likely to be subject to cultural influences. These data sources also include a set of virtually identical anchoring vignette items, a popular tool developed to alleviate the influence of response scale usage differences in cross-national comparisons. For each disability domain, three vignette questions were asked across these data sets. This study first compares disability in these countries solely based on the self-rated questions. We then examine detailed response distributions in each country and compare across countries to explore whether or not the hypothesized response scale usage differences exist. For the countries presenting this issue, anchoring vignette data will be introduced to correct for scale usage differences in both nonparametric and parametric fashion. We then compare disability status before and after using anchoring vignettes. This will allow us to examine whether and how response scale usage differences affect cross-national comparisons of disability.

Session 3 – Privacy

7. **Cultural and Interviewer Effects on Interview Privacy**
   Zeina N. Mneimneh, Michael R. Elliott, Mick P. Couper, and Steven G. Heeringa, University of Michigan; Roger Tourangeau, Westat

   Privacy (or lack of it) is an important feature of the interview mainly due to its possible effect on reporting information, especially sensitive information. Researchers rely heavily on interviewers to ensure such a setting. In reality, interviewers are essentially guests in respondents' homes and might find it difficult to achieve privacy. Thus, a substantial proportion of interviews are conducted in the presence of a third party even when the study protocol calls for it.

   Cultural factors as well as individual characteristics that are related to the three main players in the survey administration —the interviewer, the respondent, and the third party— could affect interview privacy. This presentation investigates whether there are any cultural or interviewer variations in achieving interview privacy and whether the effect of respondent and third-party characteristics on interview privacy varies by culture. To investigate such variations, survey data from 14 countries that differ in their level of affluence, family dynamics, and individuals' social power were analyzed. These 14 countries are part of the World Mental Health Surveys, a cross-national initiative that uses comparable survey design and implementation procedures. The cultural dimensions investigated include the country's level of individualism, masculinity, power distance, and wealth. Multilevel logistic regression- respondents in level 1, interviewers in level 2, and countries in level 3- was used to analyze predictors of interview privacy. Findings revealed

a significant variation across cultures and interviewers in establishing a private interview setting. The country's wealth, the country's level of individualism, and its masculinity significantly affected the privacy setting of the interview either directly or through interacting with several demographic and socioeconomic characteristics of the respondent and the third party present during the interview. The presentation concludes with practical as well as research implications related to measures of interview privacy and interviewer training protocols.

8. **Interview Privacy in Nationally-Representative Survey of Women in Qatar**
   Jill Wittrock, University of Michigan; M. Nizam Khan and Kien Trung Le, Social and Economic Survey Research Institute, Qatar University

   This paper explores interview privacy in a nationally-representative survey of women in Qatar, the 2011 Qatari Women Survey (QWS). The main objective of the survey is to evaluate changing patterns in marriage, fertility, and health among Qatari women. Rapid social change and economic development has brought changes to the traditional lives and values of Qataris, especially with regards to the life of women. The QWS discerns the degree to which Qatari women's attitudes on family formation, reproduction, and health awareness vary in the context of these transformations.

   In this presentation, we investigate the presence of a husband, father, or other adult male relative during the interview. We believe that the presence of male relative(s) is not a random occurrence and is likely to reflect cultural realities in the household. Therefore, we explore respondent and interviewer characteristics that could predict adult male presence such as respondent's age, employment status, and education, and interviewer age and nationality. We also investigate the effect of third party presence on reporting sensitive questions such as contraceptive use, a rescinded first marriage contract, and number of sexual partners. We contrast the effect on such sensitive questions with that on relatively non-sensitive ones. We also investigate whether the social distance between couples (such as age and education difference) moderate the degree to which male third party presence influences reporting on sensitive information.

9. **Respondent and Interviewer Predictors of Third Party Presence in Tunisia**
   Zeina N. Mneimneh and Julie de Jong , University of Michigan; Mansoor Moaddel, University of Maryland

   The presence of a third party during the survey interview constitutes an important contextual factor that could affect the reporting information. In spite of being instructed to conduct their interviews in a private setting, interviewers can find it difficult to establish complete privacy. Such difficulty can be more pronounced in certain cultures with specific gender hierarchy rules, high level of integration between in-group members, and lower economic levels. Still, and even in such cultures, previous research has found great variation in the privacy achieved across interviewers. While some interviewers have greater success in limiting the presence of others during the interview, others do not.

   In this presentation, we use data from a national survey in Tunisia to explore the dynamics of the interview setting and its predictors. In addition to the commonly used measures of interview privacy (who was present and duration of presence), detailed data were collected on the specific situation that lead to the presence of others during the interview. Moreover, interviewers' socio-demographic characteristics, including gender, age, education, religious attire (the presence of a veil for females), and previous interviewing experience are also

available. We first explore the dynamics of the privacy setting assembly and its variation across interviewers. Then we investigate predictors of third party presence which include interviewer characteristics, respondent characteristics, as well as their interaction.

Session 4 – Special Topics Session 1

10. **Striving for Quality, Comparability and Compliance in the European Social Survey**
    Sally Widdop, Centre for Comparative Social Surveys, City University London

    The European Social Survey (ESS) aims to achieve the highest methodological standards in all participating countries in every round. These standards concern the data collection instruments, briefing and guidance documents, fieldwork procedures and the resulting data that are released. The Core Scientific Team (CST) of the ESS and National Coordinators (NCs) in each participating country make every effort to try to ensure that the fieldwork procedures and data collected are of comparably high quality. Compliance with ESS rules, protocols and procedures is vital in order to achieve these aims.

    This presentation will provide an overview of the key requirements developed by the CST to try to ensure that fieldwork preparations, interviewer training and supervision, and data collection activities are in accordance with the requirements outlined in the ESS Project Specification and accompanying guidance documents. It will explore the difficulties experienced when attempting to ensure high quality data are collected in multiple countries in every round – e.g. related to communication, time, budget and trade-offs in decision making. Finally, it will highlight efforts to improve quality including Quality Enhancement Meetings, Field Director and NC Meetings, feedback on compliance and most recently, the identification of indicators that could be used to develop a quality profile of each participating country as well as assess overall data quality in the ESS.

    Keywords: quality control, quality assessment, compliance, comparability, equivalence, survey management

11. **Data Collection Challenges**
    Gelaye Worku and Lars Lyberg, Stockholm University

    Pennell, Harkness, Levenstein, and Quaglia (2010), Chapter 15 in the Wiley monograph, discuss a number of challenges that we face in cross-national data collection. In this paper we will check how some of these challenges have been handled by a number of ongoing 3M surveys. We will also add some further challenges, including recent findings that interviewer cheating might in fact sometimes be driven by the managers involved, that specific requirements might be ignored due to ignorance, and that the design of the data collection operation must take risks into account. We will go through the documentation of our selected 3M surveys and discuss results and implications of any studies related to the data collection operation. Finally, we will make an attempt at defining a number of design principles for the data collection operation. These principles also include thoughts on how the data collection should be controlled and evaluated on a continuing basis.

12. International Survey Data Collection Implementation: Lessons Learned
Gina-Qian Cheung and Beth-Ellen Pennell, University of Michigan

In the last several years, the Survey Research Center at the University of Michigan has partnered with four academic institutions to implement data collection projects in China, Ghana, Nepal, and the Kingdom of Saudi Arabia. In each country, interviewers carried laptop computers fitted with mobile phone cards and equipped with technical systems from the University of Michigan that were specifically adapted to each country as well as each study design. Interview data and paradata were transmitted real-time to Ann Arbor, Michigan. Daily, Ann Arbor staff work with locally-based staff in each country to monitor production and data quality.

In this presentation, we will provide an overview of our technical support model and will share the challenges that we faced and addressed during the development and implementation phases of these international projects. Specifically, we will share experiences, challenges, and solutions related to the following:

1. Balancing the standardization versus localizations of methods and approaches across different counties;
2. Making the necessary revisions to instruments or systems across multiple countries;
3. Adapting quality control protocols to country-level restrictions;
4. Building research capacity and approaches to staff and field interviewer training;
5. Adapting to country-specific hardware and software; and
6. Handling administrative and logistical challenges.

Session 5 – Response Process

13. Differential Response Styles of Subjective Life Expectancy and Cultural Differences in Time Orientation
Sunghee Lee, University of Michigan

With an unprecedented population aging pattern, the subjective life expectancy (SLE) question has emerged as an independent predictor of mortality especially for the older population. SLE asks respondents to estimate the probability that they will live up to a certain age. In spite of its predictive power for the general population, its applicability for racial/ethnic minorities in the U.S. has been questioned. This study hypothesizes that compared to non-Hispanics, Hispanics are likely to experience a higher level of difficulty in answering to SLE due to their time orientation manifested through item nonresponse and to report a 0 percent probability due to their fatalistic view on life and health. Further, we hypothesize that these response patterns are more pronounced among Hispanics interviewed in Spanish than those interviewed in English by using interview language as a proxy measure of acculturation.

Using the Health and Retirement Study, we observationally examined these item response patterns of SLE and the relationships between SLE and actuarial life expectancy (ALE) based on the Life Table and between SLE and subsequent mortality status across racial/ethnic groups. We found 1) item nonresponse and 0 probability reports were highest among Hispanics mainly due to Hispanics interviewed in Spanish, 2) SLE item nonresponse was associated with a higher mortality rate, and 3) SLE was a poor correlate of ALE and a poor predictor of mortality for Hispanics.

14. **Probing Response Processes for Self-Rated Health and Subjective Life Expectancy Questions with Monolingual English, Monolingual Spanish and Bilingual English-Spanish Speakers Using Web Surveys**
Colleen McClain, University of Michigan

Population aging is a universal pattern for virtually every country around the world. An increase in life expectancy combined with declining fertility has shifted the population age structure, and this is predicted to continue in the near future. Because population aging has profound implications for various aspects of human life, mortality prediction has become not only a relevant but also an important topic for both individuals and societies. Two survey-based mortality predictors are popularly used: self-rated health and subjective life expectancy. While the importance of these items is clear as they are empirically shown to predict true morality, little is known about how survey respondents comprehend these questions, what information they use to formulate answers and how they integrate different pieces of information. Moreover, it has not been examined whether these processes differ across cultures and if so, how they differ. In order to fill this apparent gap in the literature, we conducted Web surveys with a monolingual English, monolingual Spanish and bilingual English-Spanish speakers in the U.S. In these surveys, interview language was assigned based on their self-rated language proficiency. We experimented question formats and contexts of self-rated health and subjective life expectancy questions. These questions were individually followed by a probing question asking respondents how they arrived at the particular answer. The probing data are expected to provide insights into response processes of respective questions. This study analyzes the data from the probing questions along with answers to respective questions, socio-demographics, cultural backgrounds, language proficiency and objective health questions.

Session 6 – Paradata

15. **Using Paradata for Interviewer Data Quality Monitoring**
Nicole Kirgis, University of Michigan

On the National Survey of Family Growth, extensive use of paradata from the sample management system, organized into a production monitoring dashboard, has allowed for the conceptualization and implementation of design features that respond to survey conditions in real time, so called “responsive designs” (Groves and Heeringa, 2006). The production dashboard uses information about data collection field work to help guide alterations in field protocols during survey data collection to achieve greater efficiency and improvements in data quality.

In addition to the paradata collected by the sample management system, paradata from audit trails, the record of actions and entries within the CAPI questionnaire, are used for data quality monitoring at the interviewer level. Audit trail data include a record of every key stroke and the time spent between key strokes. Using these data, a data quality dashboard was created in order to monitor data quality at the interviewer level. Indicators include the average time spent on survey questions, the frequency of using help screens, recording remarks, checking errors, backing up in the interview, and the frequency of “don’t know” and “refuse” responses.

This presentation will discuss design and management strategies for using paradata for responsive design in survey operations to improve survey outcomes as well as discuss the implementation of the interviewer-level data quality dashboard. Examples provided will show

how this data monitoring technique has been used to identify and address interviewer data quality concerns.

16. **Using Response Time for Each Question in Quality Control on China Mental Health Survey (CMHS)**
Yan Sun and Xia Meng, Institute of Social Science Survey, Peking University

Response time for each question is audit trail data which is also one form of paradata that is generated by computer-assisted interviewing (CAI) instruments. Two approaches have been used in measuring response time in the survey literature: active time and latent time. (1) Active response time: the elapse between the interviewer finish reading a question to the respondent start to give an answer. (2) Latent response time: the moment the question appears on the interviewer's monitor to the moment the interviewer finishes coding the answer. We found that latent response time plays a critical role in quality control. In this paper we discuss the use of latent response time for each question to evaluate and improve the quality of survey data collection in China Mental Health Survey (CMHS). Specifically, we discuss how we flag suspicious samples by monitoring latent response time. We also compare the result of the interview audio record evaluation for flagged samples to the result of the not identified samples, and we find that the two groups are statistically different. We believe that using existing paradata is an efficient manner to help evaluate and improve the quality of survey data collection. Key words: response time, quality control, interview evaluation, CMHS.

17. **Using Paradata to Investigate an Unexpected Production Outcome and Associated Interviewer Behaviors**
Shonda Kruger-Ndiaye, University of Michigan

The 2013 Wave of the Panel Study of Income Dynamics (PSID) included both centralized (Lab) and decentralized (Field) telephone interviewing. Although there were no systematic differences in the sample assigned to each group, the average length of the interviews conducted centrally was significantly, unexpectedly longer than those conducted by Field interviewers from their home offices. On a study where respondent burden can impact current and future wave participation (and on-going participation of all family members), this was of great concern. However, it was unclear how interview length might correlate with interview quality and even whether the longer timings of Lab interviews was a sign of higher or lower quality or, perhaps, of differences in respondent characteristics.

Production managers used a variety of paradata tools to investigate and interpret this phenomenon. A small percentage of both centralized and decentralized interviews was digitally recorded. The outcomes of Lab and Field interview evaluations were assessed. In addition, Blaise audit trail data were parsed into an online analytical processing (OLAP) cube and the resulting dataset was used to explore the relationship between various factors and interview length, focusing on differences in interviewer behavior across the Telephone Lab and Field. Use of the audit trail data allowed quantification of differences such as number of asked questions, number of backups, frequency of remark entry, item timings and item missing data rates.

This presentation will discuss the use of paradata (specifically audit trail data) to explore an unanticipated production phenomenon. Examples provided will show how this analysis was used both during production to redirect efforts and post-production to inform future project decisions.

18. **Using Paradata to Monitor Interviewers' Behavior: A Case Study from a National Survey in the Kingdom of Saudi Arabia**
Zeina N. Mneimneh, Beth-Ellen Pennell, Yu-chieh Lin, and Jennifer Kelley, University of Michigan

The past several years has witnessed 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.

The present session will start with a general overview of paratdata and its use in interviewer-administered surveys. It will then focus on interviewers as an important source of error and the application of paradata to monitor and control this error.

The overview presentation will be followed by four cases studies on paradata use to monitor interviewer's behavior from four different surveys. 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 the most recent national mental health surveys in China and Kingdom of Saudi Arabia. Across the different surveys, pardata, specifically keystroke data and time-stamps from Blaise were used to create multiple indicators of data quality and monitor interviewer's behavior. These indicators include the average time spent on survey questions, the number of questions asked, the pauses taken by interviewers during the interview, the number of backups, and the frequency of using help screens and remark entry while administering the survey. Other types of paradata used include the maximum streak of same consecutive answers, verification results, and item missing data rates. Discussion about how paradata was used to target the review of recorded interviews and to schedule call backs will also be covered specifically for the mental health survey conducted in China.

To allow for a timely action plan, automation of the delivery and display of quality indicators to field managers is a key. Different tools could be used for this purpose. The presentations in this session will cover some of the tools used including dashboards and the OLAP Cube. Making those tools available at the beginning of data collection is extremely important when conducting cross-national surveys where time and space add to the complexity of monitoring the collection of data.

Session 7 – Special Topics Session 2

19. **The Multi-level; Multi-Source Approach to Improving Cross-National Survey Research**
Tom W. Smith, NORC at the University of Chicago

(ML-MS data). Methodologically, the use of ML-MS data in general and the augmenting of respondent-supplied information with auxiliary data (AD) in particular can notably help to both measure and reduce total survey error. In particular, AD from sample frames, databases, paradata, and other sources can be used to improve data quality and reduce total survey error. For example, it can be employed to detect and reduce nonresponse bias, to verify interviews, to validate information supplied by respondents, and in other ways. Substantively, ML-MS data can greatly expand theory-driven research such as by allowing multi-level, contextual analysis of neighborhood, community, and other aggregate-level effects and by adding in case-level data that either cannot be supplied by respondents or is not as accurate and reliable as information from AD (e.g. health information from medical records vs. recalled reports of medical care). This paper describes the ML-MS approach, considers in strengths and limitations, and indicates steps for its cross-national development and implementation.

20. Managing 3MC Surveys: Putting Total Survey Error in the Toolkit
Brad Edwards, Westat

Cross-national and cross-cultural surveys are notoriously difficult to manage. Novices may think translation is the most challenging aspect. But depending on the overall study design, many other survey elements can vary from one nation or group to another: sample design, questionnaire sections, data collection mode, survey infrastructure, schedule and cost. For general surveys, response rates are falling and costs are rising, and these trends are present in 3MC surveys as well. Clients are restless, demanding more value for less. In an era of big data, web access, and social media, it sometimes seems that probability surveys are fighting a rear guard action just to stay in the game. Survey managers – especially 3MC managers -- need help.

The Total Survey Error (TSE) framework is a tool for understanding and improving survey data quality. The TSE approach summarizes how a survey estimate may deviate from the corresponding value in the population. Measurement error, due to difficult or poorly worded items, is one reason. Nonresponse error, introduced when those who cannot be persuaded to participate in the survey are substantively different than those who do participate, is another. The framework also encourages researchers to be mindful of less-commonly studied error sources, such as coverage error, processing error and adjustment error. It highlights the relationships between errors and the ways in which efforts to reduce one type can increase another, resulting in an estimate with more total bias. For example, efforts to reduce nonresponse error may lead to poorer data quality.

TSE work has focused on the relationships between different error sources of error, on monitoring and reducing survey errors (e.g., paradata applications, adaptive/responsive survey designs accounting for multiple errors), on errors induced in combining survey data with data from other sources (e.g., imputation, administrative records), and on trade-offs between error sources in multi-mode surveys. Recently, researchers have proposed the concept of comparative error, an error type unique to 3MC surveys.

This presentation illustrates ways TSE can be applied in managing 3MC projects. Although TSE can be useful at any stage in the project life cycle, from design through data processing and analysis, this presentation will focus on 3MC survey operations. A large part of most survey budgets is allocated to data collection. TSE is a critical tool for achieving optimum overall quality with limited resources.

21. Methodological and Operational Challenges for Surveys in Multicultural and Multinational Contexts: an Integrated Total Survey Error Model
Kristen Cibelli Hibben and Beth-Ellen Pennell University of Michigan

Prepared as a chapter in the Sage Handbook of Survey Methodology (forthcoming), we present our work to date on an integrated model for the organization and discussion of error sources for surveys conducted in multicultural or multinational contexts. As background, we begin with a discussion of how multicultural and multinational surveys differ from surveys in single culture or national contexts, particularly when the goal is comparability. We also review various philosophical and theoretical approaches to comparative multicultural and multinational research with particular emphasis on current debate and research in the field of cultural psychology.

The total survey error (TSE) framework (Groves et al., 2004) is commonly viewed as the dominant paradigm for describing the statistical properties of survey estimates and organizing sources of error. Drawing on the TSE framework, we propose a model that integrates error sources encountered by researchers conducting comparative multicultural or multinational surveys. We discuss error sources – and the associated methodological and operational challenges – that are unique to or may be exacerbated for surveys in multicultural or multinational contexts. Also addressed are ethical considerations particular to multicultural or multinational surveys, as well as mixed methods, new technology and other promising approaches to survey research in this space.

Session 8 – Cross Cultural Instrument Design and Translation

22. Moving Questionnaires by Design

Peter Ph. Mohler, COMPASS and University of Mannheim and Brita Dorer, University of Mannheim

Only a few months ago members of AAPOR (American Association for Public Opinion Research) discussed a survey translation ago somewhat along the following lines: does response option translation of 3 to less than 7 not include the 7 in a German translation? Unfortunately the German version was not given. Anyway, a number of German natives as well as translation specialists rushed to help. It was quite a lively discussion on AAPORNET until one observed that using the x to less than y response form might be unnecessarily complex. Instead he offered to use the a to d; e to g; etc. option (in the specific case 3 to 6; 7 to 10; etc.). That is in a nutshell, what this paper is about: before any attempt to move an item from one nation, culture or item to another, items must be fit to move. Attempts to move unfit items leads to suboptimal translations and all to often into a laughing matter. Thus translation jumps in after an item is designed and tested for comparison, i.e. is fit to move . The paper will introduce a new terminology in an attempt to free comparative survey methods from some of the heavy bags we are still carrying along with us. The core concept proposed here is moving survey instruments being broader in scope than translation. It deliberately resonates with moving house: moving house needs careful planning or one might lose the property (three times moving is like one time burning), it also implies that moving house or home actually does almost always not mean that one transports the house and all its contents to another place, instead, it is one's core property that is movable brought into a new house or home. In this sense, moving instruments is taking them from one house into a new house with new surroundings, climate, etc. while the essential properties are sustained. Having established such a new terminology, many things just fall into their proper place such as comparative survey design as an attempt to make surveys movable all along the production line, concentrating translation on rendering measurement properties instead of satisfying language feeling of researchers, or becoming open to new insights into a subject area from the many divers cultures involved. In addition traditions that are not in line with modern cognitive survey design and modern demographics can be replaced by appropriate cutting edge methods.

23. Back Translation vs. Committee Approach Translation Experiment: An Update
Alisu Schoua-Glusberg, Research Support Services

At the 2013 CSDI Workshop we described the experiment we were conducting with an English-Polish instrument translation. We compared a translation obtained by backtranslation with one done by Committee Approach, a team approach that includes the steps in TRAPD and has gained acceptance in the industry since the mid 1990s, yet no experiments are published showing how it compares with backtranslation. For the experiment, a survey scale of 66 items was translated into Polish via Committee Approach. A comparison of a backtranslation version and the committee version were be performed to determine 1) how each process best identifies translation problems and 2) how each fares in producing an translation that native speakers in a focus group find most idiomatic. We will update workshop attendees about the experiment's findings.

24. Lost in Translation: Comparing Self-reported Health in Four Chinese Studies
Ting Yan, Mengyao Hu and Hongwei Xu, University of Michigan; Qiong Wu, Peking University

Self-rated health (SRH) is an important health measure included in many large scale surveys. Although it measures respondents subjective evaluation of their health status, it has been shown to be an important predictor of mortality and morbidity. Comparison of self-reported health in a multi-lingual, multi-culture, and multi-country context is challenging for several reasons. First, people with different cultural background have different interpretation of health. Second, people adopt different response strategies to evaluate their health and to construct their answers. Third, translation of the item may affect both understanding of the question and the response strategy to be used. The literature is rather limited with regards to the impact of translation on the answers to SRH. The translation effect has not been subject to rigid research yet. This paper takes advantage of four surveys administered in Chinese that ask about SRH using two different response scales. The question wording and the response categories are translated into different Chinese words. We examine the responses to the SRH item in these four surveys and discuss the impact of translation on the resulted answers in terms of how it affects people s understanding of the response categories and the response strategies they adopt to construct their answers. The findings of this paper will add to the survey literature on the impact of translation and will be of practical significance to researchers doing comparisons on SRH in a multi-lingual, multi-culture, and multi country context.

Session 9 – PIAAC Session 1

25. The Programme for the International Assessment of Adult Competencies: Data Collection Operational Standards
Jacquie Hogan, Westat

The Programme for the International Assessment of Adult Competencies (PIAAC) is a recently published study governed by the Organization for Economic Cooperation and Development (OECD) and is the most comprehensive international survey of adult skills ever undertaken. The survey examines literacy in the information age and assesses adult skills consistently across 24 participating countries. PIAAC focuses on what are deemed key skills for individuals to participate successfully in the economy and society of the 21st century. This multi-cycle study is a collaboration between the governments of participating countries, the OECD, and a group of international research organizations, referred to as the PIAAC Consortium.

The target population for PIAAC is a household-based sample of adults, 16 to 65 years old, who reside in each of the participating countries at the time of interview. Conducting a household-based study always presents operational challenges but doing so consistently across 24 countries is an even more daunting task. Therefore to ensure that each country followed similar operational procedures consistently the PIAAC Consortium developed a rigorous set of standards for all countries to follow during the conduct of their data collection operations. These data collection related standards addressed key activities of survey research operations including: training, interviewer quality and validation. Development of the standards was just part of PIAAC implementation process and in addition ongoing progress monitoring was conducted by the PIAAC Consortium with each country prior to and throughout the data collection period to ensure compliance and continued rigor across all study activities.

This presentation will provide an overview of the PIAAC operational standards; discuss challenges with monitoring and ensuring compliance; highlight how in some cases compromises to the standards were required to meet national operational needs and yet continued rigor was safeguarded; and the impact of the standards on the quality of the PIAAC data.

26. Benefit and Challenges in the Implementation of International Survey Design Standards in the National Context: Lessons Learned from PIAAC
Silke Martin and Anouk Zabal, GESIS

PIAAC, the Programme for the International Assessment of Adult Competencies, is an internationally conducted large-scale survey that strives to produce high quality data collected under comparable conditions throughout the participating countries. To achieve this goal a sophisticated set of standards and guidelines with regard to survey design and implementation was required and developed by an international Consortium. These standards are based on best practices in large-scale surveys and served overall as very helpful instructions and guidelines for the implementation of PIAAC in the national context. However, given national circumstances and constraints, some of these standards have raised challenges in the national implementation. The presentation will discuss the implementation and adaptation of some international standards in PIAAC Germany and will highlight both the challenges that were faced, but also some beneficial recommendations that can be drawn for future cycles or other large-scale surveys.

27. Approach to Data Quality Evaluation in an International Assessment Survey
Leyla Mohadjer, Westat

The Programme for the International Assessment of Adult Competencies (PIAAC) is a multi-cycle international programme of assessment of adult skills and competencies sponsored by the Organisation for Economic Co-operation and Development (OECD). The data collection for the first cycle of PIAAC was conducted in 2011-2012. In 2013, another group of countries began participation in a second round of cycle 1, with data collection in 2014.

Similar to other international comparative studies, the goal of PIAAC is to make inferences and comparisons across national populations on the basis of survey samples. To achieve this goal, PIAAC established an overall set of Quality Assurance and Quality Control procedures covering all aspects of the study (referred to as PIAAC Technical Standards and Guidelines (TSG). Participating countries were instructed to follow the standards to ensure the sources of survey variability were kept to a minimum and that the survey design and implementation processes of PIAAC produced high-quality and internationally comparable data.

This paper presents the results of the first round of first cycle of PIAAC focussing on the approach taken to evaluate the quality of the national data and the assessment of the fit of the data for publication purposes. The paper describes the approach used to establish quality domains and associated indicators.

Session 10 – PIAAC Session 2

28. Localization Design in PIAAC

Steve Dept, cApStAn

In PIAAC, adults from 27 countries were assessed in 26 different languages (35 country/language versions) and in two different modes: computer-based and paper-based. A combination of technical, methodological ad financial constraints called for creative solutions as regards translation and adaptation design. The initial specifications of the Item Management Portal (IMP) developed for PIAAC were written before the full implications of test delivery in two different modes "including automated scoring of computer-delivered items" were fully understood. Therefore, ad hoc solutions were developed so that stringent linguistic quality standards could be upheld and cross-linguistic equivalence could be monitored.

The tagged XML localization interchange file format (XLIFF) was used throughout the project, and the one of the PIAAC Consortium members customized an open-source translation editor to make it fit for use by countries for a double translation and reconciliation design. The hybrid design (centralised translation of the workflow, decentralised translation of the assessment materials and questionnaire modules) relied on collaborative interaction and a sense of shared ownership of the final versions. The all-important feature of the PIAAC localisation design was the centralised and standardised approach to documenting each and every step of the translation, adaptation, verification, scoring definition and final check history of each item in each language.

One of focus points in the Field Trial (FT) phase was to provide all players involved with clear and comprehensive instructions: national translation teams, IT managers, international verifiers, consortium partners. The bulk of the localisation work takes place at FT phase, and that is the key moment to achieve quality. However, since many technical and content issues were still identified during the various stages of the localization process, a comprehensive errata management process needed to be set up and a minute Main Survey (MS) verification process needed to take place: this involved a number of steps to ensure that (i) changes to the source items requested by item writers; and (ii) changes to national versions requested by country teams could be implemented correctly and consistently, and that all implications were taken into account. In tasks where the respondent needs to highlight a portion of text to indicate the response, for example, a change in wording would affect the location of the predefined blocks that are regarded as minimum correct response or as incorrect response, and therefore the automatic scoring definitions needed to be checked after edits were implemented in the stimulus.

29. **Monitoring Indications of Quality Variation in Sampling Activities in the Programme for the International Assessment of Adult Competencies**
Tom Krenzke, Westat

Monitoring indications of quality variation in sampling activities in the Programme for the International Assessment of Adult Competencies. Tom Krenzke, Leyla Mohadjer (Westat). Keywords: quality control, sample design; sample selection and monitoring. Ensuring a comparable level of quality was a primary objective in the administration of the Programme for the International Assessment of Adult Competencies (PIAAC). The PIAAC survey was conducted for the non-institutional population, resulting in about 5,000 in-person assessments in each of the 24 countries in Round 1. Quality assurance was established by the Technical Standards and Guidelines (TSGs), however, responsibilities for sampling activities were that of the country. Therefore, quality control (QC) approaches were needed to monitor activities according to their adherence to the TSGs. Several challenges occurred due to schedules, communication, variety of sample designs, skills and experience of sampling statisticians, and in the way that each country organized their sampling responsibilities. This presentation will discuss main goals of the QA and QC protocols in terms of sampling and weighting activities, and will provide some results from Round 1, as well as recommendations for future PIAAC cycles.

Session 11 – Special Topics Session 3

30. **Mixed Modes Survey Research in Cross-national Research**
Ana Villar, City University London

Survey researchers have been evaluating the consequences of adopting mixed mode designs for a number of years. The discussion has centred around whether or not mixed modes lead to higher response rates, differential measurement error and lower costs. Most of this research is evaluated in the context of surveys conducted in a single country. When cross-national survey research is involved, additional questions and concerns arise: do data quality and measurement comparability suffer when mixed mode data collection is introduced? How are response rates and sample composition in mixed mode designs compare to single mode designs? The European Social Survey (ESS) has conducted a research programme on mixed modes since its inception, in response to requests from countries to consider modes other than face-to-face for data collection that was considered too expensive and/or too challenging. In 2012-2013 a study was carried out in three countries (Sweden, Estonia and the UK) with the goal of evaluating the feasibility of collecting data in the ESS using sequential mixed mode designs. This paper will present preliminary results looking at comparability of measures within and across countries and differences in sample composition and response rates. We will also discuss lessons learned and challenges involved in conducting mixed mode research across countries where each country chose the design that felt best suited for their context. Implications of these findings towards strategic decisions about the possibility of using mixed mode research in the ESS will also be discussed.

31. Understanding Consent for Physical Measurements, Biomarker Collection, and Administrative Data Linkage in the Health and Retirement Study

Sunghee Lee, University of Michigan

The wealth of information that can be collected via contemporary surveys, including interview data, biomarker data, and consent for linkage to administrative records, provides intriguing avenues for the analysis of survey data. However, data collection beyond interview questions complicates administration as the collection of physical measures and biomarkers, as well as administrative record linkage, requires additional consent from survey respondents. Consequently, complicated errors can be introduced on such measures when the obtained data are conditional on receipt of consent. In particular, consent for some measures but not others may lead to differential nonresponse errors. These errors may be conditional on respondent and interviewer characteristics, with socio-demographic and cultural factors related to the response decision and important to consider in a comparative analysis.

The Health and Retirement Study (HRS), a longitudinal study of the elderly population, has collected physical measures and biomarker data (saliva and blood spots) since 2006 without engaging special examination facilities. Additionally, HRS solicits permission from respondents to link their survey data to the Social Security Administration (SSA) records on earnings and benefits. The study provides a unique opportunity to examine consent patterns to these requests collectively and separately in cross-sectional and longitudinal fashions. Using HRS data, we examine overall patterns of three different sets of consents: those for physical measurements, biomarker, and SSA records. Specifically, we attempt to ascertain whether consent rates are related to respondents socio-demographic characteristics (e.g., age, gender, race/ethnicity, interview language, education, income, employment status) and physical, mental and cognitive health status, their resistance to the current survey request, and their cooperation to future data collection, as well as whether the patterns differ across measures. We also incorporate detailed interviewer observations and interviewer characteristics into the analysis in order to better understand the consent patterns. In doing so, we particularly focus on the cultural background of respondents, including race, ethnicity and interview language, as well as that of interviewers.

32. Collecting Data from School Records

Kathy Buek, Mathematica Policy Research

In an evaluation of school dropout prevention interventions in primary and secondary schools in four countries in Asia, the School Dropout Prevention Pilot (SDPP) Project is using school records data to measure the impact of interventions on student dropout, performance, and attendance. The use of school records data brings certain advantages over other approaches (i.e. household surveys). However, there are a number of logistical and technical issues that must be taken into account when using records data across countries, including defining outcomes and indicators that can be used across all schools, tracking individual students within and across school years, and timing of data collection. This presentation will describe some of these issues in detail, and ways that the SDPP project is addressing such challenges. The presentation will also describe how student and teacher interviews will also be used to collect attitudinal and behavioral data not available in school records.

Session 12 – Special Topics Session 4

33. Impediments to Comparing Populations in Cross-national Surveys: Exclusion, Sampling, Noncoverage and Nonresponse
Ineke Stoop, The Netherlands Institute for Social Research/SCP

A key issue in survey methodology is representing the population. Sampling errors, coverage errors and nonresponse errors may cause the final respondents to not fully represent the population under study. In addition, social surveys tend to exclude specific difficult groups from the target population, e.g. children, the (very) elderly, non-native speakers, the homeless and the non-residential population, those who are not able to participate for physical or mental reasons, people living in far-away or hard-to reach reasons, and residents of no-go areas. In cross-national surveys the situation is even more complicated. Different sampling frames will have an impact on fieldwork and noncoverage. Differences in survey scarcity, survey attitudes and survey traditions will have an impact on nonresponse rates. Different groups will be excluded in different countries. These practices may have a larger or smaller impact depending on the survey mode. As a result, differences in outcomes between cultures or countries could be due to differences in composition of the respondents, or differences in there representing the target population, rather than substantive differences. The paper will give examples on exclusion and underrepresentation from cross-national and comparative surveys, and illustrate how cross-national differences can stand in the way of optimal comparability. Adjusting for cross-cultural or cross-national differences will not always be possible, but at least information on these differences should be available.

34. Tablets and Smartphones: Adopting New Technologies for Household Surveys in Kenya
Sarah Hughes, NORC at the University of Chicago

Tablets and Smartphones: adopting new technologies for household surveys in Kenya. The Computer-Assisted Personal Interview (CAPI) household survey has been a preferred data collection mode on studies for which paper and pencil, telephone or self-administered surveys are not appropriate or desirable and for which adequate funds are available. Given the great expense associated with face-to-face interviewing, in-person data collection is adapting rapidly to new and cheaper technologies for data capture and transmission. Moreover, new tools such as tablet computers and smartphones have expanded the field for computer-assisted data collection in recent years to developing countries. From July 2012 to March 2013, NORC conducted a CAPI household survey in 15 Kenyan cities among a sample of 14,600 households. Enumerators used tablet computers (7-inch screens with an Android platform questionnaire) to carry out a household listing of approximately 150,000 households and administer the 30-minute survey to respondents. The survey was conducted in English or Swahili, depending on the respondent's preference, and the questionnaire was programmed to allow interviewers to switch between languages as needed during the interview. Two enumerators carried out a subsample of interviews using smartphones equipped with the same software in place of their usual tablets. NORC researchers studied response data and interviewer debriefing information to assess differences in quality between tablets and smartphones. The presentation will focus on 1) the results of a comparison of the quality of data collected using smartphones and tablets, 2) the results of interviewers perceptions of their own and respondents reactions to the use of phones and tablets. We will also include a general discussion of the advantages and challenges of collecting data using smartphones and tablet computers in the context of both a multilingual environment and a developing country.

Session 13 – Harmonization

35. The (Un)intended Consequences of Comparative Data Sharing
Peter Granda, University of Michigan

Appearing in 2008, with significant revision two years later, the Cross-Cultural Survey Guidelines (CCSG) placed a series of best practices before the research community that covered the entire survey production life cycle from original design to dissemination of final data and documentation products. While intended principally for social scientists, many guidelines are relevant to disciplines in the natural and physical sciences, particularly in a world where sharing data is no longer considered simply the right thing to do but is increasingly viewed as a necessity by both principal investigators and funding agencies. This presentation will describe how the publication of the cross-cultural guidelines contributed to good practice in ways perhaps unanticipated by its authors. It will focus on one specific guideline: data harmonization. It will explore how this guideline has assisted epidemiological researchers in their efforts to pool data from many different sources into harmonized datasets to study the development and progression of chronic diseases. It will address such questions as:

- How do the harmonization methods employed by these epidemiologists agree or differ from those used by social scientists?
- How important is harmonization in this community?
- What practical harmonization work has been done in studying chronic diseases and what tools has this research community developed?
- What is statistical harmonization and how has it been used?

In addition to having the opportunity to include many more references on the CCSG web site, this development should lead other authors to see what similar (un)intended consequences their work produced or encouraged in different disciplines.

36. Survey Data Harmonization: The Issue of Data and Documentation Quality in Cross-National Surveys
Marta Kolczynska and Matthew Schoene, Ohio State University and Polish Academy of Sciences

With extensive data archives in place and further archiving efforts in progress, the focus on combining and merging of existing and available data in order to build on what had already been done seems to be the optimal strategy and logical next step of developing the research infrastructure, but " as suggested by the scarcity of harmonization projects in the social sciences - such efforts are often neither rewarded, nor incentivized." This presentation will provide an overview of the project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective" which consists in ex-post harmonization of data related to political attitudes and participation, broadly defined, from over 20 cross-national survey projects worldwide, supplemented by a linked dataset of country-level variables, and altogether covering the period 1983-2012. Differences in arguably all aspects of design and implementation between individual survey programmes, and the additional variation between countries and over time, make this an especially challenging, but at the same time instructive experiment. The presentation will focus on survey quality. Data quality assessment is an important, although often neglected element of any data analysis, and especially crucial in data harmonization projects, where the quality of data and documentation varies considerably between individual surveys. Within our project, we based the quality assessments on information provided in survey documentation, such as information about questionnaire pre-testing, translation method, sampling, response rates, item non-response to

selected questions, and presence of fieldwork control, as well as the use of comparable measures of education, income, and occupation. We will present the resulting variation in survey quality, alternative ways of evaluating quality for ex-post harmonization purposes, and implications for the harmonization process. Additionally, we would like to discuss the idea of incorporating quality measures into substantive analyses, as well as the potential for standardization of survey documentation, if not the survey process itself.