* Invited Papers

Tuesday, July 26th, 2:00 p.m. - 3:30 p.m.
Session: Adapting Interviewer Training Across Cultures

Chair: Esther Ullman, University of Michigan
Location: Michigan Ballroom I

Perspectives on international training

Katherine Mason, RTI International
Steven Litavecz, RTI International
David Plotner, RTI International

In this presentation, Ms. Mason will discuss experiences with international trainings which have been conducted in the following countries: Nigeria, Kenya, Mexico, Indonesia, Ghana, Zambia, Thailand, India and China. Topics will include cultural perspectives on time management, reading items verbatim and challenges with multiple translations, dealing with proxies, respondent privacy, challenges with subcontractors and data transmission issues.

Challenges and lessons learned of conducting computer assisted personal interviewing (CAPI) training and providing capacity building supports for a national household panel survey in Ghana

Yu-chieh (Jay) Lin, University of Michigan

The knowledge sharing of moving from paper to computer assisted personal interviewing (CAPI) and launching the national household panel survey in Ghana among global survey research and operational team members provides unique insights on adapting interviewer training with local contexts and overcoming both infrastructure and cultural challenges. Lessons learned include using the onsite train-the-trainer experience to finalize interviewer training components, customizing training activities based on trainees' immediate feedback and learning progress, being flexible with different working styles and communication approaches, understanding of differences and reacting quickly, and developing innovative technical solutions for field data collection, data sharing and analyses, and quality control. This presentation provides management and technical examples for audiences who are particularly planning to conduct survey research and data collection in developing countries or areas where need capacity building supports.

Interviewer training for a pre-school evaluation in Chad

Nathan Jones, University of Wisconsin Survey Center

Many researchers and staff providing technical support for data collection in developing countries face challenges when hiring and training local interviewers. In order to apply best practices for data collection, assure cultural adaptation of survey measures, and effectively work with multi-cultural data collection teams, interviewer training needs to be adapted to the different cultures where it occurs. The panel, Training in Developing Countries, will feature presentations by representatives from several survey organizations. This presentation will highlight lessons learned while training interviewers and conducting surveys for parents and children attending a preschool program for Darfuri refugees living in the Goz Amer refugee camp in Eastern Chad. I will focus on strategies for training inexperienced interviewers, maintaining data quality, and field work management in harsh conditions.

Tuesday, July 26th, 2:00 p.m. - 3:30 p.m.
Session: Approaches to Test for Measurement Invariance

Chair: Eldad Davidov, University of Zurich
Location: Michigan Ballroom II

Measurement invariance of different dimensions of nationalism in the ISSP : Comparisons over two time points and four countries

Peter Schmidt, University of Giessen
Jan Ciecuch, University of Zurich
Eldad Davidov, University of Zurich

Recently there has been a controversy about the use of techniques for establishing measurement invariance in the Journal Comparative Politics 2015 and 2016 dealing with scales in the World Value Study. In our study we evaluate two new techniques for establishing measurement invariance: Bayesian approximate invariance proposed by Muthen/ Aspourov(2012) and van der Schoot(2013) and the alignment procedure proposed by Aspourov/Muthen (2014). We analyze data from two waves of the ISSP identity module over several countries and evaluate the results based on the classical invariance tests compared with the bayesian approximate invariance test and the solution given by the alignment procedure.

The cross-country comparability of the immigration module in the European Social Survey 2014-15

Jan Cieciuch, University of Zurich
Eldad Davidov, University of Zurich
Peter Schmidt, University of Giessen
Rene Algesheimer, University of Zurich

A special module about attitudes toward immigration and threat due to immigration was implemented in the 7th Round of European Social Survey (ESS). In our project we set two goals. The first one was to establish in a theory-driven way latent variables based on items included in the module. These latent variables can be used by researchers in their substantive work on immigration using the ESS. The second goal was to test for measurement invariance of these scales across 15 ESS countries. We proposed the four following latent variables: allowing for immigrants belonging to different ethnic groups than the majority population into the country; qualification for entry; and two types of threat due to immigrants, realistic and symbolic. First, we tested each latent variable in each country separately in single Confirmatory Factor Analyses (CFAs). Next, we tested for measurement invariance of each latent variable using multigroup Confirmatory Factor Analysis (MGCFA). We differentiated between three levels of measurement invariance, configural, metric and scalar, and we applied two approaches: an exact and an approximate measurement invariance approach. If full or partial exact measurement invariance could not be established, we tested whether approximate invariance was given. Configural and metric invariance was supported for all constructs across most countries. Unfortunately scalar invariance was supported for the latent variables only across a subset of countries. The subset of countries where approximate scalar invariance was established was larger than the subset of countries for which exact measurement invariance could be established.

Comparing groups that are only partially and approximately comparable: an adaptive Bayesian approach

Daniel L. Oberski, Tilburg University

Muthén & Asparouhov (2012) introduced the idea of Bayesian approximate measurement invariance : rather than assume that all measurement parameters are equal across groups, a fudge factor is introduced that allows for relatively small random differences. The fudge factor must be specified in advance and results can be rather sensitive to this choice (Rudnev 2015). Moreover, when subsets of items not approximately invariant, the approximate invariance procedure does not do well at detecting these violating (partially noninvariant) items and can lead to

serious bias in the estimates of interest (Van de Schoot et al. 2014). Muthén & Asparouhov (2013) called this the alignment problem and suggested a solution based on factor rotation methods. This talk discusses a different possible solution to the alignment problem that follows naturally from the Bayesian approach and connects directly with the regularization literature (Tutz 2012; Hastie et al. 2015). Our approach is to adaptively learn from the data which items are approximately invariant, and which are not. We discuss simulation results that compare the performance of this procedure with that of the standard approximate MI model. We also apply our new approach to empirical data, demonstrating how our approach may be useful for comparing groups that are only partially, and approximately, comparable.

*** Measurement invariance in international large-scale assessments: Integrating theory and method**
Fons van de Vijver, Tilburg University, The Netherlands
Ralph Carstens, The IEA Data Processing and Research Center, Germany
Wolfram Schulz, The Australian Council for Educational Research, Australia

One of the aims of international large-scale assessments (ILSAs) of educational achievement is to collect standardized data that allow cross-national comparisons of student achievement, behaviors or attitudes, and the influences of school and classroom factors or family background on those outcomes. Complex modeling of cross-national data requires that the items designed to measure a latent factor have invariant psychometric characteristics, which means that they measure the same trait across countries, the measured latent constructs have the same meaning in all participating countries, and survey respondents interpret the items in a similar way. Non-invariant measures might due to systematic biases in the measurement instrument or differences in the way specific items are responded. Lack of measurement invariance can introduce bias and limit comparison across countries. This present study contributes to invariance evaluation over and beyond the existing investigations through the modeling of differences in measurement. We start with an overview of the basic concept of measurement invariance and its challenges with regard to ILSAs. We evaluate measurement invariance with the exact invariance assumptions and integrate a relatively new prototype of the assumptions with substantive insights about measurement bias from different factors namely in the context, item and person levels. We present a stepwise strategy for evaluating measurement invariance with a productive way of dealing with the unattainable ideal of strict invariance assumptions. With an empirical example, we argue that a generic latent structural and measurement modeling or simple structure of a common-factor model is unlikely to yield fully comparable scores within the contexts of ILSAs. These strategies could help to integrate modern, less stringent approaches, such as approximate invariance assumptions, to measurement invariance. We follow the evaluation of measurement invariance with substantive theories about the target constructs and assessment processes in our analyses, which could improve our understanding of the nature of the sources of measurement invariance. We also present an empirical example using ILSA data, and discuss some general consequences in evaluating measurement invariance across many groups, along with limitations and strategies.

Tuesday, July 26th, 2:00 p.m. - 3:30 p.m.
Session: China Household Finance Survey
Chair: Gina-Qian Cheung, University of Michigan
Location: Great Lakes E

China Household Finance Survey
Xin He, Survey and Research Center for China Household Finance
Shu Xu, Survey and Research Center for China Household Finance
Yin Zhan, Survey and Research Center for China Household Finance

Panel presentations: 1. Introduction to the China Household Finance Survey (CHFS.) The last three waves' sample designs. (8438 households in 2011; 28000 households in 2013 and over 40000 households in 2015). 2. Recruiting, Training and Managing Interviewers. In the presentation, we will share how we use our own recruiting and training

system for the tasks of on-line signing-up, resume selection, qualification verification, training information, examination, admission, in-group assignment, material management, interview monitoring, etc. 3. In this presentation, we will discuss how we designed and implemented a completed system which has the following major functions: manage interviewers, collect data, allocate samples, design questionnaire, survey and transmit back data, etc.

Tuesday, July 26th, 2:00 p.m. - 3:30 p.m.
Session: Data Collection Challenges: Case Studies 1

Chair: Stephanie Chardoul, University of Michigan
Location: Great Lakes A/B

* Data collection in cross-national and international surveys: Regional case studies

Kristen Cibelli-Hibben, University of Michigan
Beth-Ellen Pennell, University of Michigan
Jennifer Kelley, University of Michigan
Yu-chieh (Jay) Lin, University of Michigan
Sarah Hughes, Mathematica

The last couple of decades have seen continued growth in the number of surveys worldwide. This is the case for both periodic cross-national studies such as the World Gallup Poll, the Pew Global Attitudes Survey, and the Global Barometer Surveys, as well as for one-time single country studies. Results from such surveys are increasingly sought to inform public debate and scholarship particularly in transitional countries and countries facing rapid political and economic change. In this paper we examine recent developments in survey data collection, including the continued growth of multinational, multiregional, and multicultural (3MC) surveys and the proliferation of surveys generally worldwide, particularly in low resources settings and countries with relatively little survey research tradition. We also examine the increased global demand for surveys to monitor and evaluate interventions. Across these various domains is the increasingly important role of new technologies to increase data quality and speed of data access. Finally, we provide an overview of common challenges in data collection and highlight solutions that have been developed through the exploration of regional case studies. We then conclude with a discussion of future directions in international data collection.

* Data collection in cross-national and international surveys: Latin America and the Caribbean

J. Daniel Montalvo, Vanderbilt University
Mitchell A. Seligson, Vanderbilt University
Elizabeth J. Zechmeister, Vanderbilt University

Survey research in Latin America and the Caribbean, a region with a long but uneven experience in survey research (Zechmeister and Seligson 2012), is enjoying a period of major improvements in quality, scope of coverage, and data availability. Technological advances have accompanied a diffusion of professional and scientific expertise in survey methods, all serving to increase the quality of survey data collection in the region. However, improvements remain spotty, with many commercial and academic organizations having not yet upgraded to a full suite of modern scientific best practices. At the same time, comparative survey work fortunately is now beginning to include large segments of the Caribbean, which had previously been largely excluded from these studies, so that it is now possible to draw comparisons not limited to mainland Latin America.

In this chapter, we document some of the notable efforts made in advancing high quality scientific survey research in Latin America and the Caribbean, but also take note of the increasingly serious challenges that data collection in the region confronts from three trends: 1) high and increasing levels of crime and violence, 2) frequent difficulties in obtaining comprehensive high quality census data and census maps, and 3) restrictions on freedom of expression,

especially in the media, in some countries. Especially with respect to the former two issues, we discuss techniques for addressing these challenges, as well as for using new technology to greatly improve the quality of data via the reduction of fieldwork error and fabrication. Since none of these challenges and advances are inherently unique to the region in which we do most of our work, we advance this discussion with the goal of sharing techniques for overcoming or at least minimizing the difficulties imposed by these challenges to data collection both in and beyond the Latin America and Caribbean region.

*** Survey data collection in Sub-Saharan Africa (SSA): Challenges, strategies, & opportunities**
Sarah Hughes, Mathematica
Yu-chieh (Jay) Lin, University of Michigan

This presentation begins with an overview of the general setting for data collection in sub-Saharan Africa (SSA), including sampling and coverage problems, unique respondent characteristics, variation in response issues, interviewer characteristics or effects, presence of conflict zones, arduous and costly transportation to outlying regions, and the lack of communication infrastructure, etc. We then describe some unique challenges as well as opportunities emerging in data collection in SSA. With a relative lack of experimental literature on data collection in SSA, this chapter includes two case studies to illustrate common challenges and practical solutions. Where literature is scarce, the authors' observations and personal experience collecting survey data in Burkina Faso, Ghana, Kenya, Mauritania, Rwanda, Senegal, South Africa, Togo, and Uganda help fill gaps, alongside information collected in interviews with practitioners and survey methodologists working throughout SSA. We conclude with descriptions of innovations in survey research and emerging technologies that hold promise for improved statistics and data collection in SSA and suggest areas for future research.

Tuesday, July 26th, 2:00 p.m. - 3:30 p.m.
Session: Data Integration and Analysis
Chair: Irene Rioboo Leston, Eurofound
Location: Great Lakes D

Research plan for a combined analysis of the EQLS and the EU-SILC
Irene Rioboo Leston, Eurofound
Tadas Leoncikas, Eurofound

In the monitoring of living conditions and quality of life in the European Union there are two main sources: the European Quality of Life Survey (EQLS) and the European Union Statistics on Income and Living Conditions (EU-SILC). The EQLS is a questionnaire-based interview survey that covers adult (18+) population of all the EU Member States, including also the EU Candidate Countries as well as other European countries. This survey was carried out by Eurofound in 2003, 2007, 2011, with the next round taking place in 2016. The EQLS is focused on both objective conditions and subjective assessments, on both individual living conditions and societal characteristics. It contains information on subjective well-being, social exclusion, social participation, volunteering, trust, societal tensions, household composition, economic situation, work-life balance, housing and local environment. For the next wave indicators on accessibility and quality of public services are expanded, especially regarding healthcare, long-term care and childcare. The EU-SILC is an annual data source launched in 2003 and coordinated by Eurostat, based on data collected by National Statistical Institutes. Its reference population is defined as all private households and all persons aged 16 and over within the household residing in the territory of the Member States at the time of data collection. The EU-SILC provides data on income, poverty, social exclusion and living conditions in the European Union. In 2016 an ad hoc module on access to services will be run, covering child care, school and studies, life-long learning, health care, home and personal care. Both sources play a key role in the analysis of living conditions and quality of life. Therefore the possibility to combine the different aspects covered by the EQLS and the EU-SILC makes their statistical matching an extremely interesting approach. An exercise with 2007 data was carried out by Eurostat to assess the

link between both sources, concluding that the difficulties found could not guarantee the quality of the integration. The aim of our work is to develop an improved methodological plan for their matching considering the updated data available and the new modules on public services that both statistical sources include in 2016. As the microdata will not be available until 2017, this is a theoretical research with examples based on previous rounds, which presents an analysis of the metadata and a review of the main statistical methods for survey matching, providing a methodological plan to be applied when data became available.

Applying statistical matching methods for a better measurement of work-life balance in Europe

Irene Rioboo Leston, Eurofound
Agnes Parent-Thirion, Eurofound
Mathijn Wilkens, Eurofound
Greet Vermeylen, Eurofound

Eurofound, the European Foundation for the improvement of living and working conditions is carrying two household surveys: the European Quality of life surveys (EQLS) is targeting citizens aged 18 or older residents in all the EU Member States and other European countries, fieldwork for its new wave is taking place in 2016; the European Working Conditions Survey (EWCS) is a survey of workers. Its last edition took place in 2015 and covered 35 countries. Both surveys address work-life balance from a different perspective. The EQLS allows to assess work life balance from the perspective of workers and non-workers and to assess association with quality of life, whereas the EWCS allows exploring work-life balance of workers and its association with individual, household, job and organisational characteristics. Bringing the two surveys together is likely to provide new information on job quality and work life balance and quality of life. The objective of this work is to carry out the statistical matching between the EQLS and the EWCS on work-life balance, our set of target variables. This implies the integration of both surveys on the basis of the relationship among the variables they have in common, thereby providing a synthetic data set. With this exercise we cover all the methodological aspects of the integration process, which is conducted with the last updated microdata available. The exercise is based on three main steps. The first step focuses on the analysis of the metadata. This covers the comparison of the reference populations, the sample designs, the survey methods and the concepts involved. The aim is to select a potential subset of common variables that can play the role of bridges in the matching, therefore predicting the target variables. The final selection is done considering the marginal distributions of those variables. When bridging variables have been selected, in the second step, several statistical methods similar to imputation methods (correlation/association measures, regression models and cluster analysis) are applied in order to match the two datasets, thus providing a single combined dataset on work-life balance. Finally, as last step, the quality assessment of the statistical matching is performed through the evaluation of the results obtained.

Combining face-to-face interview and web add-on data in the European Quality of Life Survey

Daphne Ahrendt, Eurofound
Eszter Sandor, Eurofound

The 4th wave of the EQLS is currently in preparation, with fieldwork planned for autumn 2016. Whilst face-to-face interviewing remains the main data collection mode survey in all 33 survey countries (EU28 + 5 candidate countries), Eurofound will carry-out a follow-up web-based experiment in five EU Member States: Denmark, Germany, Poland, Spain and the UK. In these five countries with (relatively) high internet penetration rates, respondents to the main survey will be asked to complete a 15 minute on-line follow-up questionnaire. Furthermore, non-respondents will be asked to answer a very limited number of essential questions online. It is expected that approximately 20% of main EQLS respondents will complete the online follow-up questionnaire, providing 1000 internet responses in total (200 per country). To prepare for the experiment, Eurofound carried out a feasibility study in which different scenarios were evaluated. This revealed that the follow-up of the main EQLS respondents, complemented by the follow-up of the EQLS non-respondents, provides a good balance between the relatively low complexity of implementation, general applicability of findings, opportunities to compare data quality with the main EQLS, potentials for substantive

added value, and cost efficiency. An expert review of the EQLS questionnaire did not indicate inherently unsuitable questions for web administration, but a careful questionnaire design is needed to ensure high data quality and better comparability between the two modes. Larger measurement differences between face-to-face and web modes will likely occur on sensitive questions and questions prone to social desirability bias. This effect is commonly and consistently observed in surveys similar to the EQLS. However, questions with important substantive value are still recommended for inclusion in the web questionnaire despite the potential susceptibility to such an effect. The presentation will consider data-integration and comparability of the two modes in more detail. By the time of the Conference the items will have been piloted providing already some insight on the effects of social desirability of the two modes. We expect that while web responses will likely be less biased by social desirability and item non-response due to question sensitivity, the comparability with the face-to-face mode may be compromised. The session will provide an opportunity to discuss the types of challenges encountered in combining different modes. Unlike most of the research that focuses on methodological limitations, the focus is on practical solutions and on maximising the information potential of using different modes.

Tuesday, July 26th, 2:00 p.m. - 3:30 p.m.
Session: Questionnaire Design to Facilitate Translation

Chair: Brita Dorer, GESIS
Location: Huron

Improving the translatability of source questionnaires: Learnings from first-hand advance translation experience

Danuta Przepiórkowska, GESIS Leibnitz-Institute for the Social Sciences

In recent years it has been increasingly recognised that questionnaire translation performed at an early stage of a cross-national survey project can contribute to the overall survey quality. A pre-final version of a questionnaire can be translated into a small number of languages (preferably from different language groups) to check for any problems and traps which do not seem obvious at the surface but are likely to emerge in the translation process. 'Experience has shown that many translation problems linked to source text formulations only become apparent, even to experienced cross-cultural researchers, if a translation is attempted (Harkness & Schoua-Glusberg, p. 105). It is important that such an early-stage translation is performed by highly skilled translations who do not only have experience in questionnaire translation and knowledge of questionnaire design but are also capable of spotting and analysing potential problems that may be relevant for other languages and language families as well. The results of such translation-based analysis are then fed back to the questionnaire design team and can be used to eliminate errors, improve question wording and enhance the overall translatability into all survey languages. This kind of exercise has been termed 'Advance Translation' and has been conducted, for instance, since round 5 of the in the European Social Survey (see Dorer 2011). In the meantime, the methodology has been adopted by other projects as well. This paper describes the first-hand experience of the advance translation approach applied to two rounds of the European Social Survey (5th and 7th) and one round of the European Working Conditions Survey (6th). In particular, it focuses on the findings which were revealed via the translation from English (the usual drafting language in surveys) into Polish (a structurally different Slavic language). It describes the learnings that emerged from the advance translation and were shared with the ESS and EWCS questionnaire design teams. The paper also looks at the problem categories used in those three exercises, debating about their precision/vagueness as well as possible improvements and additions.

Six cost-effective modules that considerably improve a master questionnaire before the actual transadaptation process begins

Musab Hayatli, cApStAn
Manuel Souto Pico, cApStAn

In multilingual surveys, there is a strong trend towards performing more upstream quality assurance work to reduce the need for downstream corrective action. The late Professor Harkness relentlessly insisted on how important it is to craft questionnaire items carefully before they serve as a basis for adaptation into multiple languages. Her holistic approach contributed to raise awareness of questionnaire localisation issues in item writers, investigators and language professionals. In this presentation, we shall take a closer look at six cost-effective steps that can take place before the actual translation/adaptation process begins. These modules significantly enhance linguistic quality as well as cross-national, cross-linguistic and cross-cultural comparability. Module 1: An essential operation in preparing a master questionnaire for translation is parsing and segmenting the text. Parsing is about deciding what is translatable and what is not, and protecting the latter. Segmentation is about splitting the text into smaller, manageable parts (sentences or paragraphs). The optimal result of this process is that only the translatable parts of the text are editable for the linguist whereas the untranslatable parts are either hidden or locked. Module 2: It is interesting to define project-specific attributes, which can later be checked in the target versions by means of automated checks. This includes determining a maximum number of digits for keys and captions in graphic materials, defining forced, identifying the need for dynamic text, or determining what metadata needs to travel with the files, and in what form. Module 3: Before translation begins, terms and expressions are extracted from the master documents. Once the glossary is constructed according to certain rules, it is possible to set up automated checks on consistent adherence to the glossary. Module 4: Creating Style guides includes creation of and adherence to sets of typographic conventions, forms of address, notation of numbers, time, mathematical units, currencies, footnotes or endnotes, quotes, indents, etc. Module 5: To create language-specific rules requires input from linguists, so it is important to formulate the questions that need to be answered; to collect and organise responses; and then to operationalise i.e. to convert information into rules. Module 6: If concepts, alternatives, objectives, possible ambiguities or intended meaning of certain terms or expressions in the master questionnaire are explained clearly and concisely, the impact on quality is immense. That is why item by item translation and adaptation notes are an essential component of a robust translation/adaptation design.

Integrating translation and adaptation in the earliest stages of questionnaire design

Maurice Martens, CentERdata

The Translation Management Tool (TMT) is an online environment specifically designed to support questionnaire translations. It makes questionnaire texts available to translators in an intuitive way. It saves translators the effort of going through complex programming code to find and adapt texts. It displays the questionnaire in the correct order and supports text roles like help texts, interviewer instructions, answers, and even filled variables which are joined together on one screen. Once a translation is entered into this system, TMT can insert the translations back in the questionnaire source code or export to other formats that can be further integrated into questionnaire development processes. This software and working method has been originally designed for the Survey of Health, Ageing and Retirement in Europe (SHARE) , and now extended to accommodate a complex "Translation, Review, Adjudication, Pretest and Documentation" (or TRAPD) design for the European Social Survey (ESS). It is gradually supporting many more large multinational surveys. The use of this system by multiple large and influential parties should guarantee its long term availability and will make sure the system will remain robust and up-to-date according to the latest standards in translation, survey, and information technology. CentERdata participates in the European Union's Horizon 2020 research and innovation programme project SERISS, Synergies for Europe's Research Infrastructures in the Social Sciences where it implements other improvements to the TMT. This paper will present the work done on translations for multimode studies. The TMT supports not only the translations between languages, but also designed a process to support adaptation and documentation from already defined and translated existing capi and papi solutions to a cawi solution. There are many requests from translators to add tools to this system that help them with their work, like translation memories, thesauruses, dictionaries, and spell-checkers. This paper will discuss how some of these features can be integrated. CentERdata has teamed up with cApStAn to explore how to link the TMT to the various languages libraries and tools, like VeryFire and MemoryLn, they use for their verification processes. Another new feature is the generation of questionnaires directly out of the TMT, this will make it easier to verify the effects of

changes in translations, especially with dynamic texts. We will also discuss how the ever-growing underlying database of translated questionnaire items could be used to improve questionnaire development even further.

Tuesday, July 26th, 4:00 p.m. - 5:30 p.m.
Session: Benefits and Challenges of Open-ended Questions in Cross-cultural Surveys: Methodological Aspects, Use and Analysis

Chair: Cornelia Zuell and Evi Scholz, GESIS
Location: Great Lakes D

Social desirability and left-right scale placement in a cross-cultural perspective

Cornelia Zuell, GESIS
Evi Scholz, GESIS

This paper is about effects of social desirability on left-right scale placement in a cross-cultural context. Social desirability as a tendency to respond into direction of what is perceived to be the socially desired answer is an issue for survey research. Social desirability might not only distort self-reports on sensitive questions like violence, illegal acts, sexual behavior, or even income but social desirability might also affect answers wherever terms in question wording have a positive or negative connotation. While the intention for a survey researcher is to finally reduce a social desirability bias, the first step is to identify such a bias. Answers on questions on political ideology in cross-national perspective might be distorted and biased by differences in social desirability of the ideological labels due to differences in political system, history and culture. Left-right self-placement on a uni-dimensional scale is one of the standard questions in many social and political surveys to measure respondents' ideological orientation in a minimalist way. We have asked about respondents' self-placement and tested respondents' individual associations with the terms left and right by asking open-ended probe questions in an experimental online survey fielded in Canada, Denmark, Germany, Hungary, Spain, and the U.S. in 2011. We have automatically coded open-ended answers using an extensive coding scheme covering the multiple dimensions of left and right. We classified the individual categories on left and right according to three potential meanings: positive, negative and neutral or ambiguous. Respondents' aggregated assessment was used to define the socially desirable for each country separately. We tested whether ideological self-placement is influenced by social desirability, in particular if a 10-point scale is used offering no midpoint vs. if an 11-point scale was used offering a midpoint. Preliminary results seem to confirm the idea that respondents who cannot choose a scale center tend to answer into the country-specific social desirable direction.

Rules and regulations of health professions. A comparison of different institutions in EU countries

Anne Schaefer, University of Applied Sciences Fulda

Referring to the calling - Different cultures might assign different meanings and interpretation to the same term. We are interested of the response in a special topic of a specialised population to legal issues. In this present work, fact questions are the focus in our cross-cultural and cross-institutional study. Data were obtained from the International Survey. Dentists in Europe and Pharmacists in Europe . Basis of the 2013 realised mail survey are the ministries of the EU-Countries, and the Members of the Council of European Dentists, and the Members of the Pharmaceutical Group of the European Union. We analysed the extent and the cause of possible differences. Our multivariate results show that differences at the level of institutions as well as on the individual level were determined. Answers are more reliable and valid by the ministries, persons with higher work experience, higher language and writing skills, and the perceived quality of the questionnaire.

Mid-point probing and mixed mode on the LR-Scale. An application of the GESIS-Dictionnaire

Volker Hufken, Univesity of Duesseldorf

In this paper, we used a category follow up probe administered to respondent who initially select a point in the 9-point left-right scale, to determine whether they selected this alternative in order to indicate distinct opinion, or to indicate that they do not have an opinion on the issue (position). We find in the cross-section CATI- and in an online access panel survey that more than one third of responses turn out to be 'left' or 'right' and more than every fourth turn out to be 'don't know'. A reallocation of these responses from the mid-point alters the inferences. Our findings have important implications for the design and analysis of bipolar rating scales especially of the left-right political orientation scale.

Tuesday, July 26th, 4:00 p.m. - 5:30 p.m.
Session: Data Collection Challenges: Case Studies 2

Chair: Stephanie Chardoul, University of Michigan
Location: Great Lakes A/B

* Survey research in India and China

Charles Lau, RTI International
Ashish Kumar Gupta, TNS
Ellen Marks, RTI International
Chan Zhang, Fudan University

This presentation explores survey research in the world's two largest countries, India and China. The massive populations and land area of these two countries have far reaching implications for survey management, design, and implementation. At the same time, India and China are vastly different political, cultural, and economic settings. In this presentation, we highlight differences between and within these two countries. We focus on three areas: (1) Survey mode; (2) Gaining cooperation from government and local leaders; and (3) Linguistic issues. Drawing from real-world experiences, we describe challenges and solutions to common issues in survey design and implementation in these countries.

Multicultural, multilingual, multisocial, multi-everything: Empirical learning across 7 years of NIDS in South Africa

Mike Brown, Southern Africa Labour Development Research Unit

The National Income Dynamics Study (NIDS) is the first national panel study of individuals of all ages in South Africa. Some 30,000 individuals are tracked for face to face livelihood interviews along with 14,000 co-residents. The NIDS environment differs from most national panel surveys in that it faces the challenge of operating in a developing country. Currently having completed field for the fourth wave. NIDS has built up an armoury of processes and approaches to overcome the extreme diversity in peoples, landscapes and infrastructure across South Africa. This presentation is a practical summary of these learned approaches covering areas such as: Logistics difficult? Well that's just tough because a convenience sample will not be adequate. We need strategies for the seasonal effects on the landscape, dirt roads, large distances with no facilities and migrations. What respondent compensation is appropriate? Not only is much of the sample desperately poor but given South Africa's world leading inequality, a significant number are wealthy. With South Africa's political history of institutional inequality there is a heightened need to be even-handed when considering an incentive system. How to maximise the impact of using public funds in a developing country? Use of public funds in a developing country necessitates that a study is of itself developmental. Consideration must be given to developing skills to those who in South Africa are defined as from previously disadvantaged groups. What about security? Hijacking, theft, violence and attempted murder are aspects of NIDS fieldwork which must be mitigated. This influences enumerator choice, training, geographic distribution, gatekeeper relations, working practices, HR practices and more. Multiple heterogeneous gatekeepers? Deep local knowledge and systemization of this knowledge is required to gain safe and low attrition access to many areas in South Africa. Culture and race politics? Culture and race politics are a common thread through everyday life of post-apartheid

South Africa. Knowledge of this is vital but can also lead to questionable assumptions that must to be challenged. 11 official languages further complicated by dialect and differing semantics. Poor educational standards? South Africa is a country of have and have-nots across a variety of indices. NIDS has had to cope with widespread poor basic education not only in respondents but also in the available pool of enumerators. Respondent expectation of long term assistance working in developing countries means that respondents frequently get an impression that participation will mean significant improvement to their situation.

Core questionnaire and methodological criteria for working, employment and health conditions surveys in Latin America and the Caribbean

Pamela Merino-Salazar, CISAL (Center for Research in Occupational Health), Universitat Pompeu Fabra, Barcelona, Spain
Fernando G. Benavides, CISAL (Center for Research in Occupational Health), Universitat Pompeu Fabra, Barcelona, Spain
David Gimeno, University of Texas Health Science Center at Houston, School of Public Health, San Antonio Campus, USA

Morbidity, disability and mortality related to work-related injuries and illnesses are a worldwide public health concern. This concern is particularly high in low and middle-income countries like most in the Latin America and Caribbean Region, which has a critical need for the identification, prioritization and monitoring trends of these issues. In view of the need for improved information on occupational health, over the last decade, several of the countries in the region started individual efforts to implement their firsts Working, Employment and Health Conditions Surveys. Comparative research can help to identify mutual learning and cooperative problem-solving. The comparability between these surveys, however, is limited due to methodological differences, especially in the covered population, the place of interview and question wording. To improve the comparability of future surveys, an international and multidisciplinary group of 28 experts with experience in the development, implementation and analysis of Working, Employment and Health Conditions Surveys developed, through consensus, basic methodological recommendations and a core questionnaire. The consensus process was based on the surveys available in the Region (Colombia, 2007; Argentina, 2009; Chile, 2009-2010; Central America, 2011; and, Uruguay, 2012) and it was structured in two steps: (a) agreement on items to be included in the core questionnaire through two online rounds of questionnaires and one face to face meeting, where also the basic methodological recommendations were established; and, (b) definition of the questions to better measure the intended variables through six online rounds of questionnaires and one face to face meeting. The final consensus included main methodological recommendations such as conducting in-home interviews rather than workplace administered interviews and thus reaching to both the formal and informal working populations, random probability sample selection, and meeting specific criteria for conducting field work to avoid selection and information bias, such as number of visits, days and time to conduct interviews; and interviewers training. The core questionnaire comprises 77 questions organized in six dimensions: a) sociodemographic and labor characteristics, b) employment conditions, c) working conditions, d) health status, e) resources and preventive activities, and f) family characteristics. The adoption of this proposal may improve the information on work and health of the working population in Latin America and the Caribbean.

Tuesday, July 26th, 4:00 p.m. - 5:30 p.m.
Session: Global Polling: Methodological and Quality Considerations

Chair: Anita Pugliese, Gallup
Location: Great Lakes E

Sampling across diverse countries and interviewing modes

Rajesh Srinivasan, Gallup

The Gallup World Poll started in 2005 with a basic set of rules and standards around sampling of households as well as respondents within households both for face-2-face as well as telephone methodology. The expectation was that it could be applied fairly uniformly across the set of countries that were being surveyed where those methodologies were applicable. The quality of data available to implement those rules and the changing conditions on the ground have necessitated experimentation both with modes and related sampling protocols. In this paper we will discuss how sampling strategies at both household and respondent level have evolved over time across specific countries and the resulting impact it has on the quality of data.

Data quality procedures

Anita Pugliese, Gallup

Detailed data quality procedures are followed to tie everything together at the last stage. We will describe our process of data quality evaluation, which consists of several dimensions. Representativeness of the whole of country dataset is evaluated, as well as the micro level of investigating item level errors (such as an error in translation or missing response options). In additional to on-the-ground field quality control performed by our partners, we have in-house capability to view interviewer-level quality metrics. Other methods we use include looking at country trends over time, and patterns of changes on questions that are related. We compare changes in data with current events in the country; examples include, natural disasters, changes in government/leadership, changes in currency valuation.

Use of external data for validation

Dan Foy, Gallup

From official censuses to private sector indices, the demographic patterns, development indicators, economic metrics, and attitudinal data derived from World Poll data are evaluated against government indicators, NGO reports, comparable opinion research findings, and other reputable third party sources. Conducting these validations also helps identify patterns to guide subsequent analysis and modeling efforts. The panel will review some of the common data sources Gallup leverages while validating the World Poll and discuss the various challenges encountered while identifying, monitoring and handling external data sources as well as the processes for utilizing external data during construct development, quality control, and analysis.

Capacity building and quality control in authoritarian regimes

Neli Esipova, Gallup

Conducting comparable and standardized annual surveys in countries with authoritarian regimes is significant challenge. In some of these countries there is only one or no data collection company at all. With no choice for partner selection, we need to put extra effort into the development of a partner company and building this company capability. Surveying in authoritarian countries requires careful planning: from choosing the safest data collection modality to obtaining government permissions and eliminating sensitive questions, and then to hiring the right interviewers and conducting field quality control. We will discuss our experience and solutions in dealing with these situations. If people are afraid that government is tapping their phone, we cannot conduct telephone data collection even if the telephone coverage is sufficient. In the case of restricted areas in a country, it may be necessary to bring all interviewers to a non-restricted area for training and we may need to find creative ways to perform field quality control.

Tuesday, July 26th, 4:00 p.m. - 5:30 p.m.
Session: Innovative Tools in Computer-assisted Survey Measurement: Opportunities and Challenges

Chair: Silke Schneider, GESIS
Location: Michigan Ballroom I

Computer assisted measurement and coding of educational qualifications in multicultural surveys: A new set of survey tools

Silke Schneider, GESIS - Leibniz Institute for the Social Sciences

Because of complex institutional differences between educational systems across the world, educational attainment is notoriously difficult to measure in a 3MC survey context. So far, surveys have only offered measurement instruments referring to the educational system of the survey country, which is not necessarily the country the respondent was educated in. Between 2013 and 2016, GESIS has therefore developed context-sensitive tools for measuring educational attainment in cross-cultural computer-assisted surveys, e.g. surveys of migrants, or cross-national surveys. This presentation will give an overview of the project and present its rationale, which is based on 3 assumptions: 1) There is a lack of consistency and thus comparability across surveys and countries in the measurement processes and outcomes concerning the core variable of educational attainment. 2) Populations are increasingly mobile, leading to small but significant numbers of respondents with foreign qualifications amongst survey populations in many countries. Surveys of migrant populations in individual countries are a specific type of cross-cultural study that has grown recently, however without much input from the 3MC community. 3) More and more surveys are conducted using computer administrated questionnaire, whether with or without interviewer (CAPI and CAWI surveys respectively). In order to improve the measurement quality of this core survey variable, the project has thus developed a short and adaptable cross-cultural question module for educational attainment, a database of international educational qualifications including comparative coding information, as well as a search interface connecting the two during the interview process. Further presentations in this session will give more detailed information about both the question module and interface on the one hand, and the database and education coding on the other hand.

Asking about education in multinational, multiregional and multicultural contexts: Results from cognitive pretesting in two countries

Roberto Briceno-Rosas, GESIS - Leibniz Institute for the Social Sciences

Nowadays researchers from different scientific fields are looking for survey tools that provide them with a context-adaptive way of measuring respondents' educational attainment. This is possible by means of software that combines a large database of educational qualifications from each country with interfaces that allow respondents to search for their own educational qualification. Fundamental for the tool to work are the assumptions that (a) respondents understand the survey question on the highest educational qualification achieved presented to them in the same way and as intended, and (b) they are able to make use of the search interface provided in the survey environment to report their qualification(s). In this paper, I present the results of two cross-cultural cognitive studies that focus on testing the understanding of the question on educational qualifications and the usability of the CAMCES web survey (CAWI) interface. The first study was conducted in Germany with both natives and migrants. The second study was conducted in Venezuela. These studies allow us to understand how potential respondents understand the concepts used in the questionnaire, as well as the cognitive process of answering the questions. With this knowledge, we could point out the challenges for achieving equivalent measurements of educational attainment and optimize the questions and interface for future users.

Using FACS and FaceReader for analysis of nonverbal cues in respondents' emotional reactions in sociological survey

Puzanova Zhanna V., Peoples' Friendship University of Russia
Larina Tatiana I., Peoples' Friendship University of Russia
Tertyshnikova Anastasia G., Peoples' Friendship University of Russia

Sociologist receives information about the subject of his research - on the state of society, its various sectors, evaluation of social phenomena and other problems - from people. Non-verbal characteristics of human behavior are the spokesmen of human emotions, of what person wants to say, or, vice-versa, trying to conceal. Subject of non-verbal behavior is important not only in psychology, but also in sociological studies (mass surveys, focus group studies, studies of consumer behavior), but currently there is no unified analysis technology for non-verbal reactions in empirical sociological research. The article focuses on three directions , the first - how to apply the analysis of non-verbal reactions of the respondents to the questionnaire for the polls on pilot study. Within a course of earlier studies participants has been found a possible fixation and analysis of non-verbal reactions of the respondents to the questionnaire for mass surveys during the pilot studies using the method of interview (validation method). The obtained results can be used for removal and adjustment of issues that may potentially contribute to the bias of sociological data and the results of sociological research. The second direction is to study the possibility of using non-verbal reactions in the analysis of focus group studies in which non-verbal component is represented in the broadest way, because of the group dynamics that appear during discussions. Until now, there is no a unified framework for analysis of non-verbal reactions of participants. In a group discussion, the main focus should be on fixing the postures, gestures and physical location changing. The third direction is the development of guide for interviewers when dealing with interviewees, depending on their psychological type, which is set on the base of their responses to a few questions. The main emphasis of the article is using FACS (Facial action cording system - system of taxonomizing human facial movements by their appearance on the face) and FaceReader (the complete facial expression analysis software) for the above purposes.

Development of a coding platform

Maurice Martens, CentERdata

As part of the European Union's Horizon 2020 research and innovation programme project SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences), CentERdata is involved in a workpackage that includes the creation of tools that help with the coding of socio economic survey questions. The module should facilitate surveys in EU-28, OECD and almost all EU+OECD associated countries plus another 47 including: Russian, Mandarin, Arabic, Hindi and Bahasa, servicing a total of 99 countries. This task aims to program a web-based module for survey questions and answers (Q&A) for the five socio-economic variables and for the auxiliary variables required for valid coding to serve computer-assisted surveys using web, mobile, tablet, laptop or telephone; interviewers or respondents use search tree navigation or semantic matching techniques to search the relevant databases; the module facilitates single item and batch identification. To support this effort, a website and service were implemented. At this website an overview of the available codelist in various languages is presented. It is possible to download various examples and best practices, including their source codes. The backend of this website allows registered users to collect, structure and manage the lists of codes. On this same server a service is hosted that can be called to support web questionnaires. This paper discusses the various methods and variables that can be used to call this service as well give an insight into how to use this in a survey. An overview is given of the auxiliary variables and how they can be used to adapt the suggested code in the API. The website is currently being review by experts and suggestions for improvements are being collected.

The CAMCES database and its cross-national education coding system

Verena Ortmanns, GESIS - Leibniz Institute for the Social Sciences

Part of the CAMCES (Computer Assisted Measurement and Coding of Educational Qualifications in Surveys) project is the development of an international database of educational qualifications. The database can be accessed through an interface that can be integrated in the survey software and thus allows respondents (in web surveys) or interviewers (in interviewer-administered i.e. CAPI surveys) to enter the exact name of the educational qualification. Two different interfaces (dynamic text-field and the search tree) were developed and are thus set up in the database. The key units of information of the database are the educational qualifications of (almost) all European countries. The country-specific names of education certificates were identified through research on European educational systems and also extracted from the show cards of existing cross-national surveys such as the ESS, EVS, ISSP, Eurobarometer, SHARE, and WageIndicator. Those are, in the first step, systematically laid down and ordered in the database. In a second step, the country-specific educational qualifications are linked with international standard codes to facilitate the analytical understanding of each qualification and thereby cross-national comparison. UNESCO's International Standard Classification of Education (ISCED) is commonly used in the survey landscape for measuring educational attainment. Therefore the official mappings of UNESCO and, in the case of EU countries, the EU Labour Force Survey are used, both for ISCED 97 and ISCED 11 versions. Since ISCED does not provide all distinctions considered important in comparative social stratification research and official ISCED mappings are from a scientific point of view not always valid, the ESS developed its own cross-national code frame for round 5 (2010). This can be seen as an adjusted version of ISCED 11 for research purposes (instead of official monitoring and reporting). In the CAMCES project, a new meta-classification is developed from which ISCED 97, ISCED 11, and the detailed ESS education code can be derived, which will also be introduced in this presentation.

Tuesday, July 26th, 4:00 p.m. - 5:30 p.m.

Session: Interviewer Effects

Chair: Emilia Peytcheva, RTI International
Location: Michigan Ballroom II

Toward a better understanding of interviewer effects in a nationally representative survey in Tunisia

Zeina N. Mneimneh, University of Michigan
Julie de Jong, University of Michigan
Mansoor Moaddel, University of Maryland

Research has shown that interviewers can have important effects on respondent answers. Potential bias introduced by interviewer gender and religious wardrobe on related survey items is of particular concern in the gender-segregated and religious context of the Middle East and North Africa. For example, studies in the region have found that male respondents reported more egalitarian views to female interviewers (Benstead, 2013) and that interviewers wearing Islamic (rather than secular) symbols and Islamic hijab (vs. no hijab) received increased reporting of religious attitudes either directly or through an interaction with respondents characteristics (Blaydes & Gillum, 2013; Benstead, 2014; Koker, 2009; Mneimneh et al., 2015). Moreover, we have recently shown that an interviewer's own religious attitudes affected respondent's reported religious attitudes independent of interviewer religious wardrobe. The effect of an interviewer's attitudes was as large as, and sometimes larger than, the effect of the interviewer's religious wardrobe (Mneimneh et al., 2015). The literature, however, is lacking on an explanation of the mechanism of these effects. Are interviewers mirroring the attitudes of the respondents they are interviewing or are they projecting their own attitudes on the respondents? Are the effects transmitted through potential side conversations about religious topics between the respondent and the interviewer? Using recently available panel survey data from a second wave of data collected in Tunisia in 2015, this paper investigates these research questions by looking at interviewer's attitudinal measures collected before the field work and contrasting their effects with interviewer measures collected after the field work. Moreover, observational measures on side conversations related to religious and political topics were collected, allowing for investigation of the potential mediating or moderating effects on the relationship between interviewer's and respondent's attitudes.

Effects of field interviewer gender on smoking data from a global household survey on tobacco use

Jeremy Morton, Centers for Disease Control and Prevention (CDC)

Effects of interviewer characteristics on survey results have been well documented in the literature. While results appear to be mixed, interviewer effects have been found especially in cases where the survey questions are related to a demographic characteristic of the interviewer (e.g., effect of interviewer race on race-related questions). There is a widespread belief (with limited evidence) that women from certain countries/regions in the world where smoking by females is frowned upon (e.g., East Asia, Middle East), underreport their smoking behaviors because of social desirability. While biomarkers such as cotinine are a gold standard to validate self-reported tobacco use and measure misreporting, their usefulness is limited because of cost and burden. This study attempts to measure underreporting by examining the effects of interviewer gender on self-reported smoking status. We hypothesized that females might report smoking behaviors more honestly to female interviewers rather than to male interviewers because they might feel more comfortable reporting these behaviors to women. The data used for this research come from the Global Adult Tobacco Survey (GATS) for four Asian countries: China, Vietnam, Malaysia, and Kazakhstan. GATS is a nationally representative household survey designed to measure tobacco use and track key tobacco indicators. GATS has been completed in 29 countries since it was developed in 2007-2008. In some countries, gender matching of interviewer to respondent was employed because of cultural sensitivities. In the four countries investigated however, gender matching was not used. We examined the results of current smoking prevalence among women and men by the gender of the field interviewer to determine if interviewer gender had any effect on self-reported smoking. Preliminary results suggest that the smoking prevalence rates among females in China and Malaysia were different depending on the gender of the interviewer, while we found no effects on the males. Thus, gender assignment of interviewers to respondents may be a prudent approach to reduce prevarication bias and improve the accuracy of self-reported smoking information from women.

Is acquiescence an expression of social deference? Acquiescence and interviewer effects in a survey of white and ethnically diverse Latino respondents

Rachel E. Davis, University of South Carolina
Timothy P. Johnson, University of Illinois at Chicago
Sunghee Lee, University of Michigan
Chris D. Werner, University of South Carolina
Ligia I. Reyes, University of South Carolina

Interviewer effects are well documented; however, interviewer effects are inconsistent across surveys, suggesting that additional factors should be considered. Most interviewer effects research has focused on sociodemographic variables with little attention to culture. Yet, interviewers' cultural values, attitudes, and beliefs may have a powerful effect on interviewer-respondent interactions. Some interviewer cultural orientations may increase the likelihood that respondents engage in acquiescent response style (ARS), the tendency of respondents systematically agree with items with Likert-style response scales, regardless of item content. Use of ARS differs across cultural groups and, in the U.S., appears to be particularly prevalent when surveying Latinos. ARS may also be affected by interviewer-respondent social distance. Few studies have examined relationships among interviewer-respondent sociodemographic characteristics, which may be proxies for social distance. When culturally normative, respondents who perceive themselves to be of lower social standing than their interviewers may engage in ARS as an expression of social deference. This study tests this hypothesis and extends prior interviewer effects research by examining two potential influences on ARS: (1) interviewers' cultural values, attitudes, and beliefs; and (2) respondent perceptions of interviewer-respondent social distance. This presentation will present the results of a telephone survey of 400 U.S. adults from four ethnic groups: non-Latino white; Mexican American; Puerto Rican; and Cuban American. This survey will be completed by mid-January 2016. The survey assesses sociodemographic variables, acculturation, language use, and five perceived interviewer characteristics: gender; age; race/ethnicity; Latino ethnic heritage (e.g., Mexican); and educational attainment. This survey is being administered by 33 interviewers, who are completing an interviewer survey that assesses their actual sociodemographic characteristics (gender, age, race/ethnicity, Latino

ethnic heritage, education), interviewing experience, acculturation, language use, and four Latino cultural constructs that may influence their interactions with respondents: simpatico; value for sincerity; personalismo; and respect for elders. In this presentation, we will present the results of our analyses of the influence of interviewer's cultural characteristics on respondents' use of ARS, as well as analyses exploring the influence of interviewer-respondent social distance, operationalized as perceived sociodemographic similarity and dissimilarity, on ARS. Our analyses will delineate differential impacts on ARS arising from situations in which respondents perceive themselves to be of a similar, equal, or lesser social standing than their interviewers. Findings from this study will contribute to a more nuanced understanding of the dynamic role of social distance in dyadic, interviewer interactions with white and ethnically diverse Latino telephone survey respondents.

Dependence of reported height and weight on interviewers characteristics in international settings

Hayk Gyuzalyan, TNS Opinion
Elena Nikolova, European Bank for Reconstruction and Development
Francesca Dalla-Pozza, European Bank for Reconstruction and Development

Social desirability remains one of the weakest aspects of face to face data collection mode. The immediate human interaction side of the data collection ensures high quality of collected data through higher cooperation, low proportion of interrupted interviews, higher response rates compared with other modes. At the same time, the rapport built between interviewer and respondent results in stronger social desirability bias compared with other modes, especially with modes not requiring interviewer moderation, such as CAWI and postal self-completion modes. We explore the issue of social desirability in self-reported height and weight measurements. The question on reported height and weight of respondents is asked on Life in Transition Survey III conducted by TNS Opinion for the European Bank for Reconstruction and Development. Life in Transition plans to collect over 49,000 interviews in 33 post-communist and Western European countries. The project is currently in the field. We suspect that reporting height and weight may be subject to social desirability bias. In addition to the information about respondents' reported height and weight, we also collect the height and weight information from the interviewers. We explore the issue along several dimensions. First, we suspect that there may be a correlation between respondents' height and weight and interviewers' height and weight. Further, another research hypothesis is that the strength of relation between measurements of respondents and interviewers depends on the gender of respondents: reported height may be more sensitive among male respondents, and reported weight may be more sensitive among female respondents. And finally, we will also explore the impact of interviewer's gender on the correlation. The hypothesis is that the reported height and weight recorded by female interviewers will be different from those recorded by male respondents.

Tuesday, July 26th, 4:00 p.m. - 5:30 p.m.

Session: Questionnaire Translation Methods and Approaches

Chair: Danuta Przepiórkowska, University of Warsaw / GESIS
Location: Huron

Experiment for testing questionnaire translation methods in the European Social Survey (ESS): "Ask the same question" versus more adaptive approaches

Brita Dorer, GESIS-Leibninz-Institute for the Social Sciences

The European Social Survey (ESS) is a biennial academic survey fielded in 25+ countries since 2002. A British-English source questionnaire is translated into all participating language versions. The quality of the questionnaire translations is crucial for the comparability of the resulting data; therefore, the ESS has been using different methods to ensure high quality questionnaire translations since its first round: the source questionnaire is designed to minimize translatability and intercultural problems; the final translation process follows the TRAPD (Translation , Review , Adjudication , Pre-testing , Documentation) model, based on a committee or team approach. Assessment

includes an external translation 'verification' and formal checks using the Survey Quality Predictor (SQP). So far, the ESS has followed the 'Ask-the-Same-Question (ASQ)' approach: all translations should be as close as possible to the source text, ideally 'asking the same question'. Under the new European cluster project, Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS), an empirical experiment will be carried out to test this ASQ approach versus another where national teams are given more leeway to adapt rather than translate the source questionnaire into their national contexts. The exact set-up of the experiment will be decided in 2016: 30 questions will be translated independently using both methods into a small number of languages. These translations will then be administered via a webpanel also developed under SERISS. The analysis will focus primarily on equivalence (using statistical techniques such as IRT) whilst also testing the conceptual space, using correspondence analysis, or a related technique. This will be the first empirical evidence about the ASQ approach, which has been applied by several surveys assuming it allows for more equivalence, but the scientific community is lacking a proof that it does indeed produce better, i.e. more comparable and equivalent translations than by leaving national experts more room for adaptation.

Between letter and spirit: Testing survey translations with Spanish speaking respondents

Ilana Ventura, NORC at the University of Chicago
Rene Bautista, NORC at the University of Chicago
David Gleicher, NORC at the University of Chicago
Lisa Lee, NORC at the University of Chicago
Samuel C. Haffer, Office of Minority Health, U.S. Centers for Medicare & Medicaid Services
Paul Guerino, Office of Enterprise Data and Analytics, U.S. Centers for Medicare & Medicaid Services

Elderly and disabled Americans with limited English proficiency (LEP) may be particularly vulnerable to decreased access to and satisfaction with health care, and poorer health outcomes compared to those who are proficient in English. It is essential to accurately identify the LEP status and primary language of a health care recipient, in order to implement disparities identification for research as required by the Affordable Care Act and as implemented by the Department of Health and Human Services (DHHS) data collection standards in October 2011. To address these issues, NORC, under contract by the Centers for Medicare & Medicaid Services (CMS), developed and tested a new set of LEP measures for the Medicare Current Beneficiary Survey (MCBS). The MCBS is a continuous, multipurpose survey of a nationally representative sample of the Medicare population, conducted by CMS through a contract with NORC at the University of Chicago. The MCBS collects extensive information on health care use and expenditures, sources of and access to health care, and satisfaction with care. The new measures that were developed identify the LEP status and primary language of the beneficiary, preferred language for health care, and barriers to health care. The survey items were developed in English and were then translated into Spanish.

We conducted 10 cognitive interviews in English with Medicare beneficiaries whose primary language was Chinese (6), Russian (2) and Spanish (2); these interviews were conducted with the help of respondents' language assistants (family members or friends). We also conducted 18 cognitive interviews in Spanish with Medicare beneficiaries whose primary language was Spanish. The objectives of the study are twofold and focus on how translation of survey questions, whether done formally by linguistic experts or informally by language assistants, may represent a challenge for measurement equivalence across languages. We first describe the challenges experienced by linguistic experts and language assistants in adapting questions from the source language (English) to the target language (Spanish). Next, we compare answers collected in the 10 interviews conducted in English to the 18 interviews conducted in Spanish to assess whether the language in which the interview is conducted may introduce bias. As part of this examination, we present a discussion related to close translation (word-for-word) vs. adaptation, types of questions that presented difficulties in translation, as well as problems encountered when translating idiomatic expressions.

Translating tobacco survey from English to Chinese: Lessons learned

Luhua Zhao, Centers for Disease Control and Prevention (CDC)

Jeremy Morton, Centers for Disease Control and Prevention (CDC)

Introduction: As part of a concerted effort to monitor the global tobacco epidemic, the Centers for Disease Control and Prevention (CDC) collaborated with the World Health Organization and Chinese agencies on several major tobacco surveys in China. The questions used in these surveys were frequently translated from English to Chinese, and lessons were learned during this process. Methods: A core questionnaire consisting of more than 300 demographic and tobacco-related questions was translated with necessary adaptation, by a group of native Chinese-speaking public health professionals from both CDC and the Chinese agencies. The Chinese translation was back-translated to English and compared with the original English version to ensure accuracy. The quality of the translation was further investigated in a pre-test or pilot survey to identify potential issues. Results: The English questionnaire went through a full review and validation process. Although it is usually not necessary to repeat the full review process if some of the original survey content has previously been tested, a proper evaluation of the translated questions is still very important. Based on our experiences, we identified three areas, that when addressed properly, can improve the accuracy of the translation. First, important terms should be described and explained adequately; be aware that connotations from the original language may not be valid in another language. Second, subtlety in words or phrases may be lost in translation; sometimes the hidden message is hard to clarify, and not documented in dictionaries. Third, literal or word-for-word translation, while seemingly accurate, can be deceiving due to culture differences. To reduce inconsistencies caused by linguistic and cultural subtlety, it is important to conduct local focus groups to identify potential issues. The constituency of the focus group should mimic the survey target population to minimize feedback bias. Additionally, a pre-test of the survey and the subsequent examination of the results could also help identify unanticipated language and culture issues. Conclusion: Proper translation of questionnaire content requires more than the basic knowledge of the utilized languages. Mistranslation could occur due to translators' lack of linguistic and cultural understanding of specific areas. Focus groups and pre-tests are important tools to identify and reduce potential issues when translating questionnaire content.

* Preventing differences in translated survey items using survey quality prediction (SQP)

Diana Zavala-Rojas, Universitat Pompeu Fabra
Willem Saris, Universitat Pompeu Fabra
Irmtraud Gallhofer, Universitat Pompeu Fabra

Comparative surveys from design are surveys administered in multinational, multilingual and multicultural contexts thought to have the same procedures and characteristics, with the idea to match the findings in each population of study. In this type of survey research, it is assumed that by trying to keep survey features the same to the maximum extent, the data would remain comparable. Survey translation has developed best practice procedures to translate survey instruments aiming to provide the same stimuli and measurement properties in all languages. However, it is very difficult to analyse in a systematic way which translation elements or language properties matter when a questionnaire is translated. Current procedures in translation assessment do not link the quality of the translation with a formal test of comparability. Several studies have identified translation deviations as a source of non-equivalence in assessments of survey data . Unfortunately it was detected once data was collected and survey organisations had already spent a lot of resources in data collection. This paper presents a procedure that helps to prevent differences in the form of translated survey instruments using Survey Quality Prediction program (SQP). This is a procedure to foresee translation problems that could affect equivalence before data collection. Thanks to many years of research, we know to a large extent, which item characteristics are likely to affect a measurement instrument. They are also known as formal or measurement properties of a survey item. Their effects on measurement have been studied largely in the tradition of questionnaire design. Starting in 1951, Stanley Payne’s book on the art of the formulation of survey questions already considered the consequences of different question formats and answer scales. This tradition evolved and included experimental research to show how responses change between different formulations of a same concept; the cognitive processes behind a survey response and how different properties, for instance qualifiers in answer scales, affect it. We know that characteristics such as layout, question form, response scale, labelling of response options, don't know option, length of the interview, among many

others- may increase or decrease item bias and method effects. When survey questions are designed, researchers take decisions of which item features are to be chosen. Saris and Gallhofer in Design, Evaluation, and Analysis of Questionnaires for Survey Research made an inventory of those decisions (over 80) . They developed a coding scheme for this inventory to collect comprehensive information about the characteristics of a survey item and use them as predictors for measurement quality interpreted as the variance of the observed variable explained by the variable of interest. The coding scheme is incorporated in SQP program. This paper proposes to apply this coding scheme to translation evaluation. For survey questions in different languages, one can check if their characteristics are the same when the questions are coded into a same coding scheme and the codes are compared. This makes it possible to compare the characteristics independently of the languages. The procedure, explained in the paper consists in five steps: 1) introducing questions in SQP; 2) coding the source questionnaire; 3) coding a target questionnaire; 4) compare codes of measurement properties of both versions and 5) interpretation of deviations with the translation team and actions taken in the target text. The procedure has been applied in a sample of questions from the Round 5 and Round 6 of the European Social Survey, main findings show that it was useful to prevent differences in layout of direct questions in self-administered questionnaires, missing introductions, definitions and explanations, inconsistent translation in repeated formulations that should remain constant, increased complexity of the items. The findings have also opened a debate on procedures to translate labels for categories, especially qualifiers and anchors, how to solve incompatibility between the source and the target languages in the translation of bipolar/unipolar concepts and balanced/unbalanced questions. Finally, the procedure has allowed gathering systematic information on translation decisions and how translation teams solve flaws e.g. idiomatic expressions in the source questionnaire.

Wednesday, July 27th, 9:00 a.m. - 10:30 a.m.

Plenary: Comparative Survey Design and Implementation: Past, Present, and Future

Chair: Beth-Ellen Pennell, University of Michigan
Location: Great Lakes Ballroom

What's past is prologue: The origin, development, and forming of cross-national survey research
Tom W. Smith, NORC at the University of Chicago

Cross-national, survey research emerged out of and developed along with many of the seminal megatrends of the 20th century including globalization and democratization. It was also shaped in important ways by such major historical events as World War II, the advent of post-bellum collective multilateralism, and the spread and collapse of Communism.

The development of cross-national, survey research is an example of what EverettRogers calls the diffusion of innovation. Public opinion polls were created in the United States in the mid-1930s and spread to other countries. Like all diffusions, its development and trajectory was innovation specific and was both aided and hindered by the particular characteristics of survey research itself.

Its expansion was part of the more general process of globalization. Of course in the case of survey research, globalization involved considerable interaction between the global product (survey research) and the local markets and cultures. Thus, as Anthony Heath noted, "Globalization of public opinion polls has not entailed a straightforward spread of a standardized 'product' throughout the world in terms of survey conduct."
Besides being shaped by these overarching megatrends, the development of cross-national, survey research was also influenced by important, historical events. Chief among these were the impact of World War II, the advent of post-war collective multilateralism and the founding of the United Nations, and the emergence of the Cold War and the imposition of the Iron Curtain across Europe.

This paper examines 1) the emergence of cross-national, survey research including the role of early adopters, 2) the stages of expansion from the 1930s to the present, and 3) how the foundational development of comparative research is both shaping and being shaped by contemporary, survey-research methods and practices.

Why every survey is a 3MC survey

Ineke Stoop, Institute for Social Research/SCP and the European Social Survey

At present a large number of cross-national surveys provide information to measure societal and attitudinal climates, to compare countries, and to follow trends. In the recent past, the strict methodological standards that had long been employed in many national studies tended to be beyond the reach of many comparative studies (Jowell et al., 2007). Nowadays, however, cross-national studies like the European Social Survey (ESS) or PIAAC serve as role models even for national surveys.

The more than 90 thousand registered users of the ESS (March 2016) may not all be aware of the intricacies of designing and implementing a cross-national survey, pursuing the joint aims of high quality and optimal comparability. Ideally, they should be aware of the possible impact of different sampling designs, diverging response rates, varying fieldwork organizations and interviewer staff, the limited relevance of specific concepts within some cultures, and the challenge to field translations that are functionally equivalent in every language and country. Some knowledge on these issues is indispensable, however, when evaluating results and interpreting survey outcomes.

Even though relatively few survey researchers will be actively involved in the design and implementation of cross-national surveys, closer inspection shows that national surveys have quite a number of 3MC characteristics. Obviously, even within a single-country, a survey may have to be fielded in more than one language. This is in evitable in multi-language countries such as Belgium, Canada or Switzerland, but is also the only way to allow minority language or ethnic groups to participate, and thus to achieve a complete inclusive overview of societies. In national survey different response enhancing measures may have to be implemented to reach different socio-economic or cultural groups. Interviewers will have to be trained to obtain the participation of different types of respondents and guide them through the survey questions. And the questionnaire itself should be relevant, accessible and understandable for widely different groups with different skills, competences and interests. Therefore, even though cross-national comparability is not the aim of national surveys, one could state that every survey is essentially a 3MC survey.

Being involved in a 3MC study, or even using 3MC survey data, teaches invaluable lessons to those involved in single country studies. These studies are often based on national traditions, cultural norms, and proven practices. Lessons from cross-national studies can make one aware of national habits in excluding parts of the population, e.g., people living in non-residential households, of clever ways of drawing representative samples, of the risk of involving interviewers in respondent selection, of the effectiveness of respondent incentives and advance letters in different cultures and age groups, of implicit norms in seemingly neutral questions, of factors enhancing the risk of fraud, and of the vagueness of vested questions that turn out to have never been tested sufficiently.

So, experience from 3MC surveys can help survey researchers to understand the 3MC part of national surveys, and to improve national surveys in a variety of ways.

Facing the future: Opportunities and challenges for cross-national surveys: A European perspective

Rory Fitzgerald, European Social Survey, City Univeristy London

Most large scale cross-national social surveys in Europe are funded to service academia and policy makers in order to help them address key social, health, economic and environmental challenges. Europe is facing huge challenges in these areas that can be better understood and tackled using data from cross-national surveys: an economic crisis with a young generation in search of jobs, population ageing potentially straining inclusion and innovation, climate

change with its pressures to redesign energy, transport and housing patterns, just to name some of the most urgent "Grand Challenges".

In order to better harness Europe's cross-national survey resources to address these challenges, a new project – SERISS – has been launched to exploit synergies between leading European infrastructures in the social sciences. The European Social Survey (ESS), the Survey for Health Ageing and Retirement in Europe (SHARE) and the Consortium of European Social Science Data Archives (CESSDA) have come together with the Generations and Gender Programme (GGP), European Values Survey (EVS) and the WageIndicator Survey to help prepare for the future. There is also input from the World Values Survey (WVS), the International Social Survey Programme (ISSP) and the Eurofound family of surveys such as the European Quality of Life Survey (EQLS) to ensure broader discussion and impact.

Whilst the focus for the future of many single nation surveys is on response rates and internet surveys, the demands on cross-national studies are arguably rather different. For instance, using different data collection modes across countries is fraught with difficulty and response patterns differ markedly between countries. Instead the 'Synergies for Europe's Research Infrastructures in the Social Sciences' (SERISS) initiative takes a rather broader perspective in terms of preparing for the future. Specific objectives are: to better represent the European population through more coordinated sampling: to strengthen cross-national harmonization across Europe by leveraging recent advances in questionnaire design, translation and coding techniques: and to exploit the advances in software technology for cost-effective web-based interviewing, more efficient fieldwork management and in order to support new ways of collecting data. SERISS also seeks to better connect the world of research-driven social surveys with the world of process-generated administrative and social media data, and to ensure that the ethical and data protection concerns of the respondents are properly taken into account, by creating a consistent and EU-wide framework for all social surveys.

This paper will highlight examples from the project to demonstrate how Europe's academic social survey and data infrastructures are trying to advance the field and prepare for the future. In that way cross-national social surveys will play a more visible role in our society, allowing further methodological development for the future.

Wednesday, July 27th, 11:00 a.m. - 12:30 p.m.
Session: Achieving Comparability

Chair: Peter Ph. Mohler, COMPASS & Mannheim University
Location: Great Lakes D

Translation, wording, question design, survey design: An empirical exploration of different challenges to the comparability of international survey data

Michele Ernst Staehli, FORS

In comparative surveys, much effort has been invested to improve comparability of the collected data, acting on the various sources of potential errors. Prominent survey projects such as the European Social Survey aim at harmonizing as far as possible, covering domains such as unique questionnaire, unique data collection mode, harmonized sociodemographic variables, controlled sampling, contact procedures, translations and adaptations and common basis for interviewer training. Differences between countries and regions however persist: they concern - on the one side - the measurement tool (differences in translation, culture and practices of the survey agency and their interviewers, minor differences in the survey design such as the contact letters and procedures, incentives, legitimacy of the institution asking for the interview, etc.), and - on the other side - the measured 'realities' themselves, i.e. cultural and structural differences of the object, and the context in which the reality is measured, i.e. survey climate of a country or region. In some cases differences in the measurement tool simply reflect the differences in reality and thus improve comparability (we can therefore speak of good adaptations), in some cases they don't. It is however very difficult to centrally control these aspects. The efforts employed in mitigating differential sources of errors in

international surveys, the quality standards can be that high, that they compromise the participation of certain country members and even the survival of the whole project. It is therefore central to investigate in which dimensions of the quality we best should invest in future in order to optimize the comparability of the data. For this presentation we explore different sources of data, especially ESS and ISSP, with the aim of finding out what matters more for the comparability: an identical wording (respectively 'good translation'), an identical question format, or other surveys design aspects such as the way respondents are sampled and recruited, the data collection mode, the question context, etc. We focus on two simple and widely used concepts: interest in politics and satisfaction with democracy. By crossing the results from different sources, varying the dimension of their differences such as survey, mode, sampling, countries, wording, question format, language, position in questionnaire, etc., we gain a feeling of the respective importance of these different dimensions. An additional test-retest experiment, combining question versions from several surveys in a single questionnaire, allows further identifying and disentangling the different sources of the differences. First results show that, beyond wording and format, some unmeasured and often underestimated dimensions such as the way the survey is presented, or the style and content of previous questions play an important role. With this contribution, we aim at widening the discussion about comparability beyond translations to additional survey design features. We hope this will give a new impulse to the 'ask-the-same-question' versus 'ask-different-question' debate.

Calibrating cross-national panel surveys

Laura Wronski, SurveyMonkey
MIngnan Liu, SurveyMonkey

In the context of a survey, validity is the measure of how well the results obtained from one survey compare to the gold standard. Reliability is the consistency of results that one survey yields from one period to the next. For panel surveys, calibration surveys can be run to ensure that the panel meets set standards for reliability and validity. Data quality measures are drawn from the survey satisficing literature, and include straight-lining, speeding, numeric rounding, and trapping questions. Achieving valid, reliable, and high quality results is a priority for any survey, but the difficulty of doing so is compounded in surveys that make international comparisons. Beyond translation issues and problems with cultural differences, gold standard questions may vary from country to country, making reliability and validity difficult to assess. Developed countries, such as the US and UK, typically have more reliable statistical data; developing countries are likely to have spottier data available, which can weaken confidence in survey results. In this study, we run four calibration surveys in consecutive months. Each survey contains nearly identical questions (translated for each country) on topics that can be compared to known standards, such as smoking habits and political identification. The samples will be drawn from the SurveyMonkey Audience panel (in Australia, the US, and the UK) and the SurveyMonkey Global Partner Network panels (in Brazil, China, India, the US, and the UK). Our results will verify whether the panel results obtained in different months are reliable (comparable to one another), and the results overall are valid, which in this case means comparable to whatever gold standards exist for those questions in that country. We will discuss the process of selecting the calibration questions, in particular the difficulty of finding the gold standard for comparison in each country and problems with national or cultural differences. We will also describe what obstacles we encounter in fielding the survey, and any adjustments that have to be made due to the international context.

Collecting household income globally: Number and handling of missing values

Anita Pugliese, Gallup
Ken Kluch, Gallup
Dato Tsabutashvili, Gallup

Gallup has collected household income for over 150 countries through the Gallup World Poll. The greatest utility of Gallup's income data is to compare wellbeing and other social measures among income groups within a country and globally. In order to accurately compare the data across countries, Gallup transforms the income data to International Dollars using the World Bank's ICP PPP figures for consumer goods. To capture household income, the

Gallup World Poll includes an open ended income question and a closed ended income question. One of our challenges is dealing with missing income data. Globally, between 15% and 20% of respondents do not provide an answer to either item, which varies by country. Consistently, we find that female respondents are slightly less likely to provide a household income estimate; as well as those on the youngest (and sometimes oldest) end of the age distribution. In this presentation we will describe the amount of missing income data we experience across global regions/countries, characteristics of respondents with missing values, comparison of income distributions collected in the open end vs closed ended items, and our approach in handling missing data.

Wednesday, July 27th, 11:00 a.m. - 12:30 p.m.
Session: Best Practices for Panel Maintenance and Retention

Chair: Nicole Kirgis, University of Michigan
Location: Michigan Ballroom I

*** Overview: Best practices for panel maintenance and retention**

Nicole Watson, University of Melbourne
Eva Leissou, University of Michigan
Heidi Guyer, University of Michigan
Mark Wooden, University of Melbourne

While expensive to conduct, longitudinal surveys have the potential to provide rich data for analyzing the causes and consequences of change in peoples' lives. Critical to their success is the retention of panel members. High rates of sample attrition reduce precision of survey estimates, may impart bias to population estimates, and ultimately may threaten study longevity. In this presentation, we briefly review the experience of a selection of some of the world's major longitudinal surveys in retaining sample members, and discuss the key strategies used in these studies to maintain contact with sample members and encourage ongoing participation. Panel maintenance strategies are employed at different phases of the study including in the planning stage, initial contact, baseline data collection, between waves, and during each subsequent wave of data collection. These strategies are aimed at keeping in contact with sample members, especially those who are mobile, and gaining their cooperation at multiple points in time, often over a very long period. Tracking strategies include proactive measures, such as ensuring contact information is both thorough and up to date and collecting information about the likelihood of moving, as well as reactive measures that are activated when a sample member is thought to have moved. Strategies targeted towards encouraging response include the use of study branding, customizing the approach to each sample member, offering incentives, tailoring refusal conversion, developing and retaining a well-trained, well-resourced interviewer workforce, and matching interviewers to respondents.

Best practices of panel retention and tracking – International case studies from South Africa and Asia

Yu-chieh (Jay) Lin, University of Michigan
Eva Leissou, University of Michigan

Panel surveys attempt to interview the same individuals or households in order to measure change over time, regardless of the respondents' location where they were initially sampled. The main reasons individuals move are to find better housing, because of family changes (marriage, divorce, leaving parental home), or for job-related reasons (seeking or moving to a new job). These are some of the same issues panel surveys are studying; therefore, losing sample members becomes a concern because it can affect data quality and sample representativeness.

We interviewed project managers of household-based panel studies in China, Taiwan, Nepal, and South Africa, to learn about their panel maintenance practices. The strategies and frequency in which they are used vary between countries depending on the resources available to project teams as well as the country's cultural norms. Overall, a

multi-step approach is found to be necessary for successful outcomes, including steps during the field period and between waves of data collection.

Strategies of panel maintenance and retention in the Survey of Health, Aging and Retirement in Europe (SHARE)

Frederic Malter, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Munich
Michael Bergmann, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Munich
Felizia Hanemann, Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Munich

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel study generating person- and household level data on health, socio-economic status and social and family networks. Across its six waves currently released to the scientific community it contains approximately 123,000 individuals (with more than 293,000 interviews) from 20 European countries (+Israel) aged 50 or older.

This presentation will highlight strategies employed to keeping respondents in the panel during the long tenure of SHARE and by highlighting harmonized procedures as well as many country-specific approaches necessary to obtain good outcomes. We will demonstrate respondent incentives used across countries and waves, and our PR materials we use to stay in touch with respondents. Recently we improved our web-based presence geared towards respondents and will point out highlights of these efforts during the presentation. Another aspect of our presentation will be the showcasing of those functions of our internationally harmonized software tools that enable us to track respondents over time (such as stable address, proxy contact information etc.). Finally, we will outline measures we use at interviewer trainings and in interviewer payment to maximize success at gaining cooperation of panel households/individuals. The presentation will close out with a careful assessment of the success of these various strategies in obtaining high panel retention.

Sample attrition in the China Family Panel Study

Yan Sun, Institute of Social Science Survey, Peking University, China

This paper provides an assessment of the nature of attrition in the CFPS sample between wave 1 to wave 3. More specifically, the paper presents summary statistics on the levels of response achieved in each wave, reports on observable differences between wave 1 respondents who did not respond in the follow up survey and those that did, and examines briefly the impact of attrition on population estimates.

Wednesday, July 27th, 11:00 a.m. - 12:30 p.m.
Session: Mode Effects in Social Surveys in Developing World

Chair: Curtis Cobb, Facebook
Location: Huron

The challenges and solutions to conducting computer assisted telephone interviewing (CATI) in Africa: A view from the frontline

Jain Dharmendra, TNS
Mariam Fagbemi, TNS
Melissa Baker, TNS

The challenges and solutions to conducting Computer Assisted Telephone Interviewing (CATI) in Africa: a view from the frontline. In sub-Saharan Africa, face-to-face (F2F) interviewing methods have prevailed over other methods such

as Computer Assisted Telephone Interviewing (CATI), online and mobile-based methods which were, until recently, very restricted by low penetration of telephones and internet. However, ever increasing penetration of mobile phones has opened the door for innovation in primary research practices. Within this context CATI offers one option and proffers a wide range of potential benefits in terms of speed, low cost, quality and researching difficult-to-reach populations. This paper examines the operational aspects and challenges of CATI including sampling, response rate, data collection processes, fieldwork monitoring and management, quality assurance, cultural and gender issues, and interviewer effects in conducting phone surveys in Africa. It uses the Nigeria CATI Centre as a case study. The centre has conducted over 75,000 successful interviews over two years, cutting across interviews with enterprises and with individuals and for both one-off and continuous, a wide range of thematic areas. This paper will use evidence from the researchers working on these studies, examining the following questions: 1) Response rates are a very important determinant of sample representativeness. What are the key influences on response rates in a Nigeria context and how can we maximize response rates? 2) Nigeria and many other sub-Saharan African countries are multi-lingual, with Nigeria having more than 200 languages. Specifically what is the influence of language of the interviewer and culture to response rates and data accuracy? What is the best way of conducting research in this multi-lingual environment? 3) What practical challenges are faced and how are these overcome? Overall, CATI is slowly gaining its place and becoming preferred methodology especially for a wide range of research goals and as such we would like to share practical information in this paper to optimize use and effectiveness.

Cross-cultural variation in mode effects between smartphone and computer-based web surveys

Curtiss Cobb, Facebook

Mobile devices are quickly become the dominant mode for accessing the Internet around the world. Experts predict that mobile Internet usage will overtake desktop Internet use worldwide this year (mobi-Thinking 2014). With the rapid increase in use of mobile devices for Internet, web surveys completed using smartphones and tablets have followed suit, with current estimates ranging from 7% -30% of web surveys taken on a mobile device (Maritz 2013). Survey methodologists are busy exploring the mode effects between smartphone and computer-based web surveys and developing best practices for the design of multi-mode web surveys, but they do so based on research done almost exclusively on Americans or Western Europeans. It is unclear whether the same mode effects are present to the same extent among in emerging markets, especially those where mobile is often the first and primary mode of Internet access. This talk will present a series of findings on cross-cultural variation in mode effects (both selection and measurement effects) between smartphone and computer-based web surveys using data from more than 80 countries from respondents using both computer-based and mobile devices.

Telephone sampling in the MENA: Issues and advancements

Carsten Broich, Sample Solutions

CATI surveys are a widely used method for data collection as an alternative to conventional face-to-face studies. Especially in areas where face-to-face data collection is difficult to conduct due to cultural, financial, logistical or political issues, CATI surveys remain a predominant method of collecting data. In Northern America and many Western European countries population studies have been replaced by online studies. Nevertheless due to cultural, political and financial aspects, CATI surveys remain a very favorable method to collect data in the MENA region. In order to reduce the total survey error it is important that the telephone sampling frame is setup correctly. While in Western Europe and North America most of the numbering plans are well documented, this is not the case in most of the MENA countries. To prevent undercoverage error, the sampling plan should include all allocated numbering blocks while at the same time remove non-allocated blocks to increase the working number rate. Furthermore the amount of allocated numbers within a numbering block remain very low resulting in a very inefficient telephone sample in case of no further filtering being applied. A correct categorization of status codes of phone numbers before and after data collection is also necessary in order to estimate the non-response bias. This paper will analyze research studies from recent years in which landline and cell phone numbers have been pre-filtered. Also methods for pre-screening of landline and mobile phone numbers will be outlined which can further increase the efficiency during

fieldwork. On the other hand it is mandatory that working numbers are not removed by chance. Methods of analysis to prevent this are discussed. With decreasing landline phone penetration and increasing cell phone penetration it will be required to make use of a higher share of cell phone sample. Currently most researchers are hesitant to increase the share of mobile phone penetration due to the fact that no location information is available, in contract to the US where an exchange code even for cell phone numbers exists. This paper will further analyze ways to find estimates for the location of cell phone numbers within predetermined regions.

The validity of collecting data using short-message service (SMS): Experimental evidence from four African countries

Charles Lau, RTI International
Ansie Lombaard, TNS
Lisa Thalji, RTI
Melissa Baker, TNS
Joseph Eyerman, RTI

Short-message service (SMS, or text messaging) surveys offer an inexpensive and rapid new data collection mode. SMS surveys are especially attractive in developing countries, where mobile phones are widespread and there are fewer regulations. However, the validity of SMS surveys is not well understood. Coverage errors may arise due to illiteracy or a lack of mobile phones. Non-response errors may emerge, particularly among lower income people , who disproportionately have poor phone signals, problems charging phones, and financial concerns about participating. Measurement errors are also a concern due to short questions and small phone screen sizes. However, we are unaware of any research that investigates these issues. Our study investigates whether SMS surveys can produce representative samples and accurate measurements. We are conducting SMS surveys with 4,000 adults in Kenya, Nigeria, Ghana, and Uganda. Surveys asks questions about socioeconomic status and technology use, and are based on a stratified random sample of mobile phone numbers in each country. To investigate how survey design affects non-response errors, we randomize the survey length (16 questions, 8 questions), incentive ($0.50, $1.50), don't know responses (included, excluded), and the day/time of invitation. Our analysis addresses three main questions. First, how representative are SMS surveys? In each country, we compare the SMS sample composition against Census data and two gold standard, in-person surveys. We also consider how survey design (survey length, incentive, don't know, invitation day/time) affect non-response errors. Second, we explore how the common practice of sharing SIM cards and having multiple SIM cards affects sampling weights and survey statistics. Third, we investigate measurement quality. We study primacy effects by analyzing select all that apply questions with a randomized order of response options. We also compare the levels of don't know responses between SMS surveys and our gold standard in-person surveys.

Wednesday, July 27th, 11:00 a.m. - 12:30 p.m.

Session: Nonresponse

Chair: Ineke Stoop, Institute for Social Research/SCP and the European Social Survey
Location: Great Lakes E

*** Geographic and linguistic correlates of household, adult, and minor nonresponse in California: The dual role of place and personal cultural characteristics**

Matt Jans, UCLA Center for Health Policy Research
Kevin McLaughlin, UCLA Center for Health Policy Research
Joseph Viana, UCLA Center for Health Policy Research
David Grant, UCLA Center for Health Policy Research
Royce Park, UCLA Center for Health Policy Research
Ninez A Ponce, UCLA Center for Health Policy Research

Nonresponse is a complex, multiply-determined phenomenon driven by forces at multiple levels of sample units' ecosystems. For surveys of multinational, multicultural, and multilingual populations, social and cultural norms are an important consideration. Dimensions of culture such as power distance, individualism v. collectivism, and femininity v. masculinity have been found to influence unit nonresponse. Yet other demographic summaries of geographic areas show additional predictive ability beyond cultural dimensions. Each sample unit's contact attempt history (i.e., call-specific dispositions recorded in paradata) provides detailed information about the household's likelihood of response, and community-level measures of survey difficulty, such as the U.S. Census Bureau's Low Response Score (LRS) may further help with understanding the overall nonresponse problem. This chapter uses community-level measures of cultural orientation, acculturation-relevant demographics, and survey resistance, and sample-unit-level measures of household resistance/commitment to understand the geographical clustering of nonresponse in a RDD survey of California's residential household population. Data from the 2011-2014 California Health Interview Survey (CHIS) are combined with auxiliary data from the American Community Survey (ACS) and decennial U.S. Census to create community-level measures. The predictive ability of community-level cultural dimensions is compared to call history measures of resistance/commitment, community-level survey resistance (i.e., LRS), and acculturation-relevant community-level demographics. Community-level cultural dimensions predict screener and adult response, but their effects disappear when considering other community-level factors. The LRS is an informative predictor even when considering other community-level characteristics, but it predicts only screener response, not adult response. Overall, we find that acculturation-relevant community-level demographics are more predictive of response than theoretically-derived cultural orientations. Specifically, the population concentration of low-education residents and Spanish speakers with limited English proficiency predicts greater screener response, while the concentration of immigrants predicts lower response. Once screened, response by a randomly-selected adult is higher in areas with a greater concentration of older adults, and lower in areas with a greater concentration of immigrants. Finally, measures of sample unit resistance and difficulties remain influential in the presence of other community-level information, emphasizing the importance of proximal measures. Language problems at the screener predict higher response, perhaps due to adjustments made by the interviewing staff to accommodate the language barrier (e.g., providing an interviewer who speaks the contacted person's language). However, language problems at the adult interview request predicted lower response. Implications for survey practice and theory of survey nonresponse in culturally-, demographically-, and linguistically-diverse populations are discussed.

* Does providing additional languages improve representativeness?

Michael Ochsner, FORS and ETH Zurich
Oliver Lipps, FORS

People who do not speak (one of) the survey language(s) as their mother tongue tend to be underrepresented in social science surveys. There is as a consequence some evidence that a wider range of survey languages can have positive effects in reducing representation bias. However, the decision if an additional language is offered at all and if yes, which additional language is chosen, most often follow logistic and financial reasons. Therefore, more knowledge on a) the conditions when the addition of a language may have a positive effect and b) techniques to assess the potential of additional languages is needed. This presentation will provide recommendations and the methodology for this end by focusing on who can potentially be included by offering an additional survey language and whether adding a survey language improves representativeness. For a comprehensive analysis of the role of the survey language(s) on representativeness, we use data from three general population surveys in Switzerland. We find evidence for a higher probability to lose foreigners and especially those from a country where none of the survey languages are spoken during all steps of nonobservation. Using data from the Swiss census, a potential analysis reveals that adding a language improves coverage of foreigners depending on the language added. However, with regard to representation bias we find evidence for a more complex pattern: Reduction of bias not only depends on the language chosen but also on the level of language mastery that is demanded by the survey.

* Comparing nonresponse and nonresponse biases cross-nationally

James Wagner, University of Michigan

Ineke Stoop, Institute for Social Research/SCP and the European Social Survey

There are several reasons why studying nonresponse in a cross-national context is important (see also Stoop et al., 2010). In cross-national surveys the primary aim is usually to compare countries. Differential nonresponse biases across countries may bias these comparisons. Therefore, understanding differences in response rates and the composition of response is an important first step in any comparative analysis that plants to account for potential nonresponse biases. For example, ff the noncontact rate in some countries is low and high in others, this may have an impact on the comparability of estimates related to contactability, such as time spent in outdoor activities. Factors that vary from country to country may be important in explaining differences in nonresponse bias across countries. For example, the number of requests to participate in surveys will differ across countries. In countries where surveys are rather new, higher survey cooperation may be expected. Survey scarcity may therefore have a negative impact on comparability. Empirical evidence has begun to be assembled on the potential impact of nonresponse bias on cross-national comparisons. In a cross-national analysis of nonresponse in the Labour Force Survey, De Leeuw and De Heer (2002, pp. 52-53) found that noncontact rate variation was associated with differences between countries in average household size and the percentage of young children, whilst differences in refusal rates were associated with economic indicators (unemployment rate, inflation rate). Different countries have distinct survey attitudes and survey traditions, and ease of contact and reluctance to cooperate differ between cultures. Johnson et al. (2002, p. 68) suggest that social participation patterns, socioeconomic opportunities and communication styles may influence cultural values and survey nonresponse. Stoop and colleagues (2010) explore these various influences on patterns of response in the ESS. Improving comparability with respect to nonresponse is thus a challenge. A persistent problem in the pursuit of optimal comparability in cross-national surveys is that it cannot - and is not expected to - always be achieved through the wholesale adoption of precisely the same methods or procedures across different countries. On the contrary, it may sometimes be preferable to facilitate variation in certain procedures precisely in order to achieve comparable outputs. To obtain high response rates and minimise nonresponse bias, knowledge of and adaptation to national factors is required. Sometimes differences will have to be accepted, such as the availability of different sampling frames, internet penetration or the absence of experienced interviewers. Also, different types of incentives could be used in different countries (or maybe none in some countries), different types of advance letters and different appeals to national authorities or exciting topics. Incentives will be standard in some countries and quite unusual in others. In the latter case, field organisations are reticent in handing out incentives, either because they do not think it will have the intended effect or because they are afraid it will spoil the market •. In the European Social Survey in some countries it is highlighted in the advance latter that this is a European survey. This is a perk in new EU countries. In other countries, the word Europe is carefully suppressed because of the “not unjustified“ fear that mentioning Europe will seriously put off respondents. Studying nonresponse in a cross-national context is exciting, exactly because there are no standard solutions, and best practices in some countries might be less adequate in others. Cross-national surveys, where national circumstances differ, can therefore be seen as a test bed or laboratory for nonresponse research. In practice, though, it turns out that many factors behind survey response are confounded within each country. In the European Social Survey, for instance, fieldwork is conducted by National Statistical Institutes in the Nordic countries, where interviewers are employees and the sample is drawn from the population register. Other countries use private organisations to conduct fieldwork, draw samples from address registers, and interviewers “who have to select respondents within households“ are paid by the hour. One way to keep track of all these differences is by carefully documenting the fieldwork. Contact history information or contact forms are increasingly used in many countries. With the information from these it is possible to explore differences in both fieldwork practices across countries (are calls made during the evening?) and fieldwork outcomes (how likely is it that someone opens the door at a morning call). The next step is to assess bias across countries and adjust for this by weighting. The ability of these adjustment strategies to reduce the nonresponse bias of estimates may vary by country. The number of auxiliary variables that can be used for assessing and adjusting for bias will differ across countries, partly due to use of different types of sampling frames. The associations of these auxiliary variables with survey outcome variables may vary as well. Beyond differences in the available auxiliaries, there are choices to be made the kind of adjustment strategy to use. Using the same weighting model in every country may not produce comparable quality of estimates, however, selecting the best • model in each country may produce inconsistencies of

other kinds. This chapter explores how differences with nonresponse across countries may lead to different nonresponse biases and how these biases may be addressed with the explicit goal of providing comparable estimates.

Survey respondents' poor cooperation during the Interview in the United States and Korea
Jibum Kim, Sungkyunkwan University (Korea)
Jaesok Son, NORC at the University of Chicago
Sori Kim, Sungkyunkwan University
Hee-Choon Shin, NCHS
Jeong-han Kang, Yonsei University
Jungeun Lee, Sungkyunkwan University

Do the restless or hostile respondents during survey interviews have the same characteristics across cultures? Using the 2002-2014 General Social Survey (GSS) in the U.S. and the 2003-2013 KGSS in Korea, we examine the correlates of uncooperative respondents. Interviewers assessed the respondents' attitudes toward the interviews, which range from very cooperative to hostile (not cooperative at all in Korea). By collapsing two positive and two negative categories, we create a dummy variable (uncooperative respondents=1). The percentage of uncooperative respondents range from 2.5% to 4.2% in the U.S. and from 7.7% to 16.7% in Korea. Using the logistic regression analysis, we find the following common characteristics among uncooperative respondents in both countries: older, non-married, and reluctant to give income information. Gender does not matter in both countries. The uncooperative respondents in the US tend to be non-whites and political independents, and the educational level has mixed effects. Religiosity does not make a significant difference in the US whereas less religious people tend to be less cooperative in Korea. We discuss our findings in light of the increasing usage of paradata.

Wednesday, July 27th, 11:00 a.m. - 12:30 p.m.
Session: Privacy and Ethical Considerations
Chair: Patricia Goerman, U.S. Census Bureau
Location: Michigan Ballroom II

Third-party presence in multi-national interviews: Determinants and implications in SHARE
Galit Gordoni, The Hebrew University in Jerusalem
Howard Litwin, The Hebrew University in Jerusalem
Noam Damri, The Hebrew University in Jerusalem

The present paper focuses on respondents who are subject to privacy violation in the survey interview, and it examines the implications of such non-privacy for data quality of sensitive questions. Sensitive questions can decrease a respondent's willingness to give reliable and valid responses due to the perceived negative implications of true reporting. The design requirement for privacy in surveys is intended, therefore, to prevent misreporting because others were present in the interview. However, privacy during the interview is dependent on the willingness of household members to adhere to this request. Previous cross-survey comparisons show that despite the design requirement for privacy, there is often significant presence of third parties during the interview. The current study focused on elderly respondents in SHARE, a panel survey of older adults in Europe (and Israel) - A population that is currently gaining in academic and policy interest. The rising concern is in response to the accelerated aging of the population in the last several decades, which is occurring due to increased longevity and reduced fertility. Important research infrastructures, such as the Health and Retirement Study (HRS), the English Longitudinal Study of Aging (ELSA) and the Survey of Health Aging and Retirement in Europe (SHARE), are now producing extensive longitudinal data on persons aged 50 and older, data that facilitate the study of a wide range of life aspects among older people. However, lack of knowledge on the determinants of third-person presence in such large scale surveys and on the implications of third-party presence for data quality in this particular population is evident. This lack is all the more

significant due to the potential vulnerability of older adults to the effects of power-relations in the household which come into play during the interview, and the resulting possible bias in survey data. Using data on respondents aged 65 and older from the 5th wave of SHARE, we find that third-person presence is more likely among older adults who are vulnerable, in terms of lack of health and social resources, as well as among those participating in the panel for the first time. However, the results do not support social desirability reporting on the majority of subjective well-being measures in the SHARE data (depressive symptoms, quality of life and life satisfaction). That is, the presence of a third-party did not bias or otherwise compromise the responses to sensitive questions.

Variation in third party presence during face-to-face interviews: Measurement property and interviewer effects

Zeina N. Mneimneh, University of Michigan
Colleen McClain, University of Michigan
Yu-chieh (Jay) Lin, University of Michigan

The presence of a third party in face-to-face interviews constitutes an important contextual factor that affects the interviewee's responses to culturally sensitive questions (Aquilino, 1997; Casterline and Chidambaram, 1984; Mneimneh et al., 2015; Pollner and Adams, 1994). In cross-cultural surveys, variation in third party presence during the interview adds a layer of complexity to the comparability of the data collected (Mneimneh, 2012). Two potential sources of variation in interview privacy across and within samples are the measurement property of interview privacy observations and differences in the perception and attitudes of interview privacy across interviewers. This presentation explores both of these factors for the first time using data from the Saudi National Mental Health Survey (SNHMS). The SNMHS is the first national mental health survey of the Kingdom of Saudi Arabia and is part of the cross-national World Mental Health (WMH) initiative. Interviews are conducted face-to-face using a computer assisted personal interviewing (CAPI) mode with an audio computer assisted self-Interview (A-CASI) component for a number of sections judged to be of sensitive nature to the Saudi community. Interviewers are required to record their observations regarding the presence of a third person at the end of several questionnaire sections throughout the interview, in addition to recording this information about the overall presence of a third person at the conclusion of the interview (the commonly used method of collecting interview privacy data). We use these two types of observations to discuss the potential magnitude of measurement error arising when comparing the most commonly used overall summary measure against a newly constructed section-specific measure. We then compare predictors of interview privacy for each of these types of observations using a series of multilevel models and focusing on the effect of interviewer-level characteristics and more specifically interviewer attitudes towards privacy (while controlling for respondent and household level characteristics). This is the first study to investigate interviewer attitudes towards interview privacy as predictors of third party presence. This is quite important in cross-cultural survey research given that the perception of privacy as a right to the individual (especially among in-group members such as family members) could greatly vary across different cultures such as collectivist vs. individualist cultures.

* Ethical considerations in the total survey error context

Julie de Jong, University of Michigan

Survey research inherently involves interaction with human subjects. Ethical considerations and standards guiding researchers' engagement with human subjects (i.e., respondents) have evolved over time and as survey research expands to heretofore inaccessible populations, new ethical considerations have arisen. This is especially true in the multinational, multiregional, and multicultural (3MC) context, where research often includes populations with more limited survey research tradition and where socio-political factors engender vulnerable populations in need of rigorous ethical protections. However, while a standard ethical framework exists, both legal and professional regulations as well as socio-political contexts can lead to differences in how ethical protocols are realized across countries. The interpretation and implementation of any one study design decision based on ethical considerations in an individual country can contribute to the sources of total survey error (TSE) in that country. In the 3MC context, where obtaining comparable survey data is a central objective, differences in the implementation of ethical standards

can contribute significantly to the sources of total survey error (TSE) differentially across all study countries, and contribute to the overall comparison error. The goal of researchers in a 3MC survey is to acknowledge that the application of ethical protocols may differ across study countries and consider the effects on TSE and subsequent comparison error. This chapter considers existing ethical standards and emerging ethical considerations in survey research within the TSE framework and how differences in implementation of ethical protocols in the 3MC context may contribute to comparison error.

Wednesday, July 27th, 11:00 a.m. - 12:30 p.m.
Session: TRAPD and Translation Assessment

Chair: Dorothée Behr, GESIS
Location: Great Lakes A/B

Set-up of the TRAPD model on the 4th European Quality of Life Survey (EQLS)

Daphne Ahrendt, Eurofound
Tadas Leoncikas, Eurofound

Our paper will present the set-up of the TRAPD model on the 4th European Quality of Life Survey (EQLS). The survey is currently in preparation, with fieldwork planned for autumn 2016. Part of the preparation involves developing a final source questionnaire in English, which will subsequently be fielded in 33 European countries (EU28 and 5 Candidate Countries). In total the questionnaire will be translated into around 40 language versions. As part of the questionnaire finalisation process, 30 cognitive interviews in English will be held and the draft source English questionnaire will be subjected to a translatability assessment in one Roman (French or Italian), one Germanic (Dutch or German) and one Slavic (Czech or Polish). The results of these two pre-translation exercises will be discussed in a meeting between Eurofound staff, the researchers responsible for the cognitive tests and the linguists in charge of advance translation. The EQLS TRAPD model can be outlined as follows. Two independent translations are carried out. For existing questions, for which translations are available from previous waves, will not need to be retranslated, translators will ensure coherence between the translations of the new and the existing questions by pointing out the key elements requiring attention. Should translators be of the opinion that coherence can only be ensured by changing an existing question, the suggested change (and accompanying argumentation) will to be provided to Eurofound for approval. These two independent translations will be reviewed in a meeting between the two independent translators and the adjudicator. In an interactive session the two independent translations are discussed and the final translation is agreed. The adjudicator is responsible for the final decisions about the translations. The process and outcomes of decision making about the translation of each questionnaire item will be recorded systematically, distinguishing between situations where (1) translations do not differ, (2) translations differ substantively and (3) translations differ technically. By the time of the conference the translated versions of the questionnaire will have been piloted in all survey countries. Next to the detailed discussion of EQLS TRAPD model procedures, the presentation will highlight particular questions, items and themes in the area of measuring quality of life that are susceptible for translation errors and interpretation challenges in the context of survey translator work in Europe. The session will provide an opportunity to discuss lessons learned for achieving optimal equivalence with its measurement instruments.

Implementing the TRAPD model: Team/committee approaches in practice

Alisú Schoua-Glusberg, Research Support Services

In 2002, Janet Harkness defined the TRAPD Model for survey instrument translation as consisting of several steps: translation-review-adjudication-pretesting-documentation. The strength of the TRAPD model over traditional translation and independent review lies in the group discussions and group dynamics, and where different actors bring their expertise to the table, to review, assess, and improve upon the initial draft translation.

In the decade following the initial dissemination of the model, practitioners began implementing it in survey translation around the world. For example, the European Social Survey played a major role in disseminating the use of the model across over 20 countries.

In this presentation we will discuss how the model is being implemented in different ways and contexts, and will discuss which modalities preserve the basic elements of Harkness' model and which others do not take advantage of its strengths.

Double revision approach: An efficient translation method

Oscar Riviere, TNS opinion

Over the last ten years, at TNS opinion we have developed a translation unit to cover our translation needs for the international surveys that we conduct. We have handled the translations of such projects as the Eurobarometer, EQLS, ENPI, the World Bank's Enterprise Survey to name a few. We have also our own team of over 400 professional translators covering about 60 languages. We have extensive experience of the two most common translation procedures of questionnaires: back translations or adjudication processes. However, we were often confronted with requirements from clients where these two methods were not applicable. Most often because for financial reasons or for the time needed to complete the whole process. So we had to develop a different approach that would still fulfill the quality requirements of the translations but at a lower cost and in faster way. We combined our experience of translating questionnaires for our regular clients and the standards of the language industry and developed the double revision method. The basic approach is that the translation of the questionnaire is revised by two different translators. However, we have managed to integrate some of the elements of the adjudication method. The final reviser acting as the adjudicator; all the feedback, suggested changes or comments are documented in one single document. By documenting all the steps and making them available for the adjudicator during the final revision we provide all the necessary information for the translator to finalise the local version of the questionnaire. So every translation is viewed by three translators. It is also very easy to include a manager from a local institute in the process if necessary. Further, we have defined a specific format for the translation file in Excel which allows us to give specific instructions, contextualisations, comments, etc. for each and every item. The main idea it that the translator should have all the information at disposal in the same file he/she is working in. The format for the double revision approach allows the translator to use CAT tool (Computer Assisted Translations) such as Trados or WordFast. The double revision method is very well adapted for short to medium length questionnaires that needs to go on field quickly. For instance, a 20 minute questionnaire can be translated in 20-30 languages in less than 10 days. Obviously, it can be used for very long questionnaires as well.

Assessing translations: How different checking procedures compare under field conditions

Steve Schwarzer, Pew Research Center
Katie Simmons, Pew Research Center
Gijs van Houten, Pew Research Center

Achieving a good translation of a source questionnaire from one language into multiple other languages that maintains consistent measurement properties is often a challenge. Until recently, the most commonly used method for assessing questionnaire translations was back-translation. Research initiated by Harkness's work has shown the weaknesses of this approach. This led to the development of team and committee focused approaches, which are now widely used in multi-country projects. But, these types of approaches are rather complex, costly and have implications for the timeline of projects. Translation verification is another approach that has been discussed as an alternative to back-translation procedures and committee approaches. An independent translator reviews the original translation and discusses with the original translator how to modify the translation. In our paper, we outline the extent to which the verification approach resulted in changes to translation. We evaluate whether changes suggested through the verification approach to existing translations – that were back-checked in previous editions of the survey – produce different results in terms of item non-response and in terms of the distribution of the

substantive answers. To do so, we will analyze the results of the 2014 and 2015 editions of the Global Attitudes survey, comparing changes over time between items for which the translation was changed following the verification, items for which the translation was flagged but not changed and items for which the translation was not flagged. Although the evaluation does not allow us to assess the effectiveness of either approach in an absolute sense, it should provide some insight into any relative improvements made by applying the verification approach.

Wednesday, July 27th, 2:00 p.m. - 3:30 p.m.

Session: 3MC Cases Studies

Chair: Rob Bijl, The Netherlands Institute for Social Research/SCP
Location: Michigan Ballroom II

Job insecurity and individual well-being: Moderation by labour market policies across Europe. A multilevel analysis in a cross-country perspective

Beatrice Piccoli, University of Leuven (KUL), Belgium
Hans De Witte, University of Leuven (KUL), Belgium

This is a multidisciplinary study which combines work psychology and social policy research with a cross-national comparative approach based on data from ESS and EUROSTAT of 2010, among 21 European countries. The aim is to identify potential mitigating factors of the job insecurity-well-being relationship at the country-level. Conservation of resources (COR) theory, in particular, allows us to take these different contexts into account. It provides an integrative framework for testing which contextual and individual resources help job insecure employees to maintain well-being, or to buffer the negative effect on it. Following a multilevel perspective suggested by COR theory, how employees respond to insecurity depends not only on individual-level factors but also on the country-institutional context. Our assumption is that national-level policies aimed at reducing the negative consequences related to unemployment can provide individuals with more resources to deal with job insecurity. Specifically, countries with generous Labour Market Policies (LMPs) can reduce the detrimental outcomes of job insecurity, through increasing individuals' employability (active LMPs) or protecting their income during unemployment (passive LMPs). Generous policy support is thus expected to buffer the negative consequences of job insecurity in terms of well-being. Based on hierarchical linear modelling testing cross-level interactions, initial support for this hypothesis is found. Specifically, in countries with more generous LMPs the negative relationship between job insecurity and well-being is less strong than in countries with weaker LMPs.

A comparative study of the relationship between brand trust and customer loyalty across countries, languages, customer types, and various demographic groups

Daniela Yu, Gallup

With recent scandals of Volkswagen and Baidu breaking their brand promises, brand trust becomes a hot topic for media discussion. In fact, in the modern age of increasingly popular e-commerce, brand trust is even more important for business exchanges since online transactions involve more uncertainties for the customer compared with offline such as store visits. Some extent literature has found positive relationship between trust and customer loyalty; however, few have studied whether the relationship exists in multiple contexts and have made comparisons of the strength of the relationship across multiple contexts. This study is aimed to explore whether and to what extent the positive linkage between trust and customer loyalty exists across multiple contexts defined by countries and languages worldwide, and across customer types (i.e. between B2B and B2C customers). Within U.S., the strength of the relationship was compared across different demographic groups defined by gender, age cohorts, education and geographic regions and industries. The customer loyalty includes 3 self-reported metrics, including customer satisfaction, repurchase intention and likelihood to recommend. The data was drawn from a large customer database with 14 million customers from 60 countries and 39 languages. The unit of analysis is respondents, and multiple logistic regression was applied to study the strength of the relationship. As a result, the study further demonstrated

the strong positive linkage between trust and customer loyalty in all studied groups such as countries, languages, B2B vs. B2C customers, and across different demographic groups. In addition, significant difference of the strength of the relationship was found among some groups but not others. For example, building trust is more effective to increase customer loyalty for women than men, for millennials than older counterparts, and for B2B than B2C customers. However, no significant difference exists across geographic regions within U.S. Finally, the implications and limitations of these findings are also discussed.

Public sector media convergence as a strategy for intercultural communication: A case study from the European Union

Michael Elliott, Institute of European, Russian, and Eurasian Studies at Carleton University

Can a pan-European mass communication policy and it's entrenchment in EU legislation promote and sustain an emerging culturally and ethnically diverse European political community? Media systems in Europe are largely asymmetrical to the Europeanization project due to highly nationalized media networks that are entrenched in legislation at the member-state level. The member-state fragmentation of media policy is a highly problematic notion. The French media are largely representative of French republicanism: citing freedom of expression as an inalienable right. Mutual suspicion between Muslim migrants and French citizens are underscored by inaccurate media representations of Muslims in the French press. This paper advances the solution of a converged European mass media as a tool of multicultural democracy. The European Commission has advanced several regulatory policy proposals which espouse the idea of a pan-European media network, which will include the integration of minority media channels into a larger European framework, and the creation of pluralist European electorates through the visibility of opinion leaders. An example of these opinion leaders is the large number of French Muslim soccer players who play on the French national team. Their potential role and function as mediating opinion leaders can be solidified through a policy of cross-promotional circulation and oversight of Muslim-French media at the European level. A pan-European media policy could mandate the creation of European level news networks which present informed coverage of both Muslim and French issues as European issues. For the purposes of this paper, I will premise my analysis of the European mass media as a model of multiculturalism. One of the ways to do this is to look at the dimensions of horizontal integration of European policy domains in the areas of culture, media, and telecommunication. The closer integration of these policy domains fits into the paradigm of European Integration, while satisfying the relationship between EU elites and their watchdog auditing organizations which sponsor campaigns aimed at strengthening democratic initiatives. The horizontal policy integration, in this sense, will be used as a method of investigating contemporary democratic theorists, such as Jurgen Habermas, who idealize the European Union as a public sphere bearing the potential of intercultural communication in an ethnically diverse society. Using the French-Muslim example within the context of horizontal policy integration will provide an empirical framework by which to measure the capability of engaging intercultural expressions within the theoretical accounts of the mass media as a European public sphere.

Understanding race/ethnic differences in public opinion surveys towards the Affordable Care Act

Tianshu Zhao, University of Illinois at Chicago
Timothy P. Johnson, University of Illinois at Chicago

The Patient Protection and Affordable Care Act (PPACA), commonly called the Affordable Care Act (ACA) or Obamacare, was signed by President Obama on March 23, 2010. Obamacare brought a series of significant changes to the U.S. Health care system, especially intended to reduce the number of uninsured people. However, although approximately half of the people who are uninsured are white, a greater proportion of minorities are uninsured (Teitelbaum and Wilensky, 2013, Essentials of Health Policy and Law. Burlington, MA: Jones & Bartlett Learning, p.52). Thus, race/ethnic differences in the health care system is an additional public policy issue of concern. Since 2010, Obamacare has generated intense public debate and controversy. In this project, we propose to investigate public support for and opposition to Obamacare across time using a series of 35 publically available public opinion surveys conducted since 2010 by the Kaiser Family Foundation, which include questions concerned with Obamacare. These

data sets will be pooled and their sample weights adjusted in order to examine temporal changes in support, including the effects of key events related to the implementation of Obamacare, such as Supreme Court rulings. This pooled data set will include approximately 36,500 individual-level records, which will additionally be used to investigate race/ethnic and other sociodemographic variability in Obamacare support, along with the effects of various types of respondent health insurance status. Hierarchical linear modeling will be employed to conduct these analyses. These models will enable us to control for the temporal clustering of responses and examine, using interaction terms, and variability in the Obamacare support trajectories of various population subgroups over a six-year period. The large sample size that will be available in these pooled data will provide opportunities to examine the unique trajectories of relatively small population groups, such as those of individuals with specific types of public vs. private vs. no health insurance. Our study will borrow theoretical and analytic perspectives from survey research and public policy. Findings will help inform our knowledge regarding patterns of citizen support, over time, for this landmark public health policy initiative.

Effect of smog on reports of well-being in China: A survey in Beijing and Shanghai
Chan Zhang, Fudan University, Shanghai, China

Air pollution has become a severe problem in China. So far, most of the research has focused on the health risks posed by polluted air. Little attention has been given to psychological impact of living in a smoggy atmosphere. Schwarz and Clore (1983) found that weather could affect people's judgments of well-being through its influence on mood. Through the same mechanism, air quality is likely to have similar, or perhaps even more fundamental impact on people's mood and their evaluations of the quality of their lives. This study investigates this by conducting telephone surveys of Beijing and Shanghai residents under four conditions: sunny days with good air quality (Air Quality Index 300), rainy days with good air quality (AQI 300). The data collection will be carried out in the spring of 2016. The sample will be randomly assigned to be called on one of these four types of days. The survey includes questions to measure people's mood at the time of being interviewed and their satisfactions with life overall and with different life domains. This study will assess the variation in people's mood and their evaluations of well-beings under different conditions.

Wednesday, July 27th, 2:00 p.m. - 3:30 p.m.
Session: Harmonization, Data Documentation and Dissemination
Chair: Peter Granda, Univeristy of Michigan
Location: Great Lakes E

*** The past, present and future of statistical weights in cross-national Surveys: Implications for survey data harmonization**
Marcin Zielinkski, University of Warsaw
Przemek Powalko, Polish Academy of Sciences

*** Identification of processing errors in cross-national surveys**
Olena Oleksiyenko, Polish Academy of Sciences
Ilona Wysmulek, Polish Academy of Sciences
Anastas Vangeli, Polish Academy of Sciences

This paper discusses one particular and often overlooked aspect of survey quality – the consistency between different sources of documentation containing metadata that define variables and their values in questionnaires and codebooks, and between the numerical data records. In the general survey quality framework, our analyses relate to both the Total Survey Error paradigm, and to the user-specified dimensions of quality, which emphasize, among others, accessibility and usability of the survey data. We conducted our analyses within the Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling project (funded by the

Polish National Science Centre, grant number 2012/06/M/HS6/00322). We selected seven target variables of substantive interest for the project. For these seven target variables we identified a total of 688 source variables from the 89 survey waves. For each of the source variables we cross-checked metadata from codebooks, questionnaires, and SPSS dictionaries with information in the numerical data records (i.e. data files). After cross-checking the different sources of information, we identified 154 processing errors. We propose a typology of processing errors in survey documentation. Then, we present the analysis of the prevalence of different types of processing errors, and compare their occurrence in different source variables. We finish with the general evaluation of the 89 survey waves with regards to survey quality as reflected in the degree of consistency between documentation and data records.

* Survey data harmonization and the quality of data documentation in cross-national surveys

Marta Kolczynska, The Ohio State University
Matthew Schoene, The Ohio State University

For ex-post harmonization projects that aim to combine information from different surveys into a merged dataset with common variables suitable for comparative analyses, availability of, and access to, comprehensive documentation of the source data is crucial. Despite significant improvements in the field, fully documented surveys are rare. In this paper, 22 international survey projects selected for ex-post harmonization are examined through the descriptive documentation that accompanies each survey. We identify Sample and sampling, Response rate, Translation of the survey instrument (questionnaire), Pretesting, and Fieldwork control, as key elements the survey documentation should account for, given these steps' relevance for gathering high quality data. For 1721 surveys stemming from the 22 international projects, we create dichotomous indicators recording the presence or absence of a particular type of information relating to these stages. We use them to evaluate between-survey differences in documentation outcomes, as well as differences between the international projects to which surveys belong. We argue that the rigor with which surveys report on key steps of the survey process represents, in the Total Survey Quality framework, one aspect of survey quality. Overall, we find a general positive trend over time in the quality of surveys as reflected in documentation. It shows that (a) within long-lasting survey projects, proper description of the survey process receives increasing importance, and (b) newer projects have higher documentation standards than older ones. This presentation is part of the project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective," a joint endeavor of the Institute of Philosophy and Sociology at the Polish Academy of Sciences and The Ohio State University supported by the [Polish] National Science Center (2012/06/M/HS6/00322).

* Item-specific metadata in ex-post harmonization of international survey projects

Marta Kolczynska, The Ohio State University
Kazimierz M. Slomczynski, The Ohio State University

The emergence and development of cross-national surveys in the last 50 years have greatly facilitated empirical research on a wide variety of social, economic, political and demographic processes that require a comparative approach. At the same time, repeated measurements have allowed to trace changes over time. Still, in practice, research opportunities remain to a large extent limited by the scope of a single survey program and very few research projects combine data from different surveys to extend the geographical or time coverage of analysis. This happens for many reasons, among which the most important are differences in methodologies employed in the survey process which make joint analyses of data from different projects not immediately possible. These differences can occur on the level of survey programs, such as question wording or response coding; on the level of survey waves within a single survey program, e.g. order of questions in questionnaire; or on the level of countries within a single survey round, like sampling method, survey mode, or quality control measures. So far there exist no standard and recognized procedures of dealing with such data issues allowing to combine two or more datasets. Meanwhile, the analysis of data obtained by merging several cross-national survey projects would make it possible to answer new questions and verify hypotheses that until now could not be tested. The project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National

Perspective", a joint endeavor of the Institute of Philosophy and Sociology, Polish Academy of Sciences, and The Ohio State University, aims to address issues of comparability and consists in ex-post harmonization of data related to political attitudes and participation, broadly defined, from over 20 cross-national survey projects worldwide. With regard to data analysis, we propose a new methodology of conducting cross-national research by incorporating survey metadata as variables in statistical analyses of substantive problems using harmonized survey data. These additional variables are of three kinds. The first group comprises variables referring to variables that have been harmonized, for example the type of original response scale in case of attitudinal items or the time period mentioned in retrospective questions, with the aim of correcting for harmonization effects. The second group includes information on survey design, e.g. type of sampling or survey mode. Finally, the third group includes survey quality indicators, such as response bias, the type of quality control, presence of pre-testing, or description of sampling. A synthetic measure of quality could be used to weight data from different surveys according to their quality. From the design point of view, survey data can be thought of as hierarchical structures where individual responses are the lowest of four levels: the highest level is the survey program, divided into survey waves (typically spanning from 1 to 5 years), which are made up of single-country surveys (i.e. surveys carried out in a single country within a given wave of a particular survey program). Because values of the aforementioned variables are specific to different levels: survey, survey-wave, or survey-wave-country, it is necessary to employ multi-level modeling to account for the non-independence of individual observations. In this way surveys would be used as contexts for individual-level data, an approach, which "to our knowledge - has not been used in extant research. In this paper we present the benefits of including metadata in substantive analyses, using the example of popular survey items on political trust and protest behavior as illustrations. Additionally, in discussing challenges of cross-national research, we emphasize the need for standardization of survey documentation and the survey process itself.

Wednesday, July 27th, 2:00 p.m. - 3:30 p.m.
Session: Quality Control and Monitoring

Chair: Lars Lyberg, Stockholm University
Location: Huron

* Interviewer monitoring in the Saudi National Mental Health Survey

Zeina N. Mneimneh, University of Michigan
Yu-chieh (Jay) Lin, University of Michigan

The Saudi National Mental Health Survey (SNMHS) is part of the World Mental Health cross-national initiative. SNMHS is based on a national multistage area probability sample of 5000 households. Within each selected eligible household, a male and female Saudi between the ages of 15 and 65 are selected randomly. Interviews with selected respondents are gender-matched and are conducted face-to-face using the Saudi version of the Composite International Diagnostic Interview (CIDI 3.0). Computer Assisted Personal Interview is used with audio-assisted components for sections asking about sensitive information. Respondents are also consented to give saliva samples at the end of the interview. Collected data are sent to a central server in Ann Arbor, Michigan. Interviewers send and receive data daily using a University of Michigan in-house sample management system.

This presentation focuses on the quality control procedures used to monitor interviewers collecting data in the SNMHS; and how the more traditional procedures (verification and field observation) are supplemented by a data-driven approach. The data driven approach compiles real-time substantive data and paradata by interviewer and flags interviewers requiring additional quality control assessment based on a set of quality control indicators. The quality indicators chosen for the SNMHS are classified into two groups: single occurrence indicators, and cumulated indicators. Within each of these two groups, indicators are further classified by the potential error type they could be associated with: measurement, coverage and nonresponse error. Indicators are compiled and displayed in a series of tables that are reviewed by the management team. Details on the tools and the process used to flag interviewers for

additional quality control assessment are presented. The presentation concludes with suggestions on how to improve the quality control process and make it more efficient.

* Interviewer monitoring in the European Social Survey

Lars Lyberg, Stockholm University

The European Social Survey (ESS) is a cross-national survey that aims to measure attitudinal change over time. The survey has an extensive methodological, substantive and administrative infra-structure, which makes considerable contributions to the state of the art of comparability surveys possible. Data are collected using face-to-face mode. Mixed-mode alternatives have been examined for 10+ years but the current view is that any deviations from single face-to-face mode would compromise comparability without being considerably less expensive.

Thus, interviewers collect the data. A number of quality assurance techniques are used to make this process as good as possible. The central ESS team has developed training materials that can be adapted to local circumstances and implemented by national coordinators and national data collection organizations. Organizations should hire experienced interviewers used to probability sampling. They are briefed about ESS specifications and the justification for them. For instance, there is an upper limit on interviewer workload set at 48, since we know that workload size is correlated with an increased design effect. Ignoring interviewer effects leads to an overestimation of the relationship between variables and an underestimation of standard errors. This phenomenon is not well known in all camps. For instance, according to Beullens and Loosveldt (forthcoming) all 221 ESS substantive research articles published in 2013 ignored information about interviewer variance and its effects.

Monitoring in ESS is done through contact forms and back-checking to see if interview has taken place. There is little information about the actual outcome of these interviewer controls. The results come in late and damage might already have been done. To shed light on data quality we have to rely on a wealth of evaluation studies on interviewer-related nonresponse and measurement issues. Here we describe in more detail what ESS has in place regarding interviewer errors and how interviewer quality assurance and quality control can be improved and become more timely.

* The Consumer Pyramids Survey

Mahesh Vyas, Centre for Monitoring Indian Economy
Dhananjay Bal Sathe, Centre for Monitoring Indian Economy

Field execution of household surveys face several serious challenges that can raise questions regarding the accuracy of the data collected. CMIE's Consumer Pyramids Household Survey addresses these challenges by inducting technological solutions into processes that are designed specifically to overcome these challenges on three counts. First, it ensures that the interviewer does reach the household that is to be interviewed. Second, it ensures that the data collected is based on an interview and finally, that the data collected by the interviewer are consistent and likely to be correct.

All the data collected is checked for all of the above by a team that is independent of the execution of the survey, in real-time -- ie as the data is collected. Discrepancies observed are relayed to execution team in real-time to ensure immediate and realistic resolution.

The approach is based on the optimal use of real-time information. Every member of the execution team has access to a little more information than the members they supervise. Knowledge that the overall system is continuously monitored, that the management has additional information, and that it uses this information to verify current processes, motivates the team to adhere to the survey protocols.

*** Television audience measurement panel in India**
Sharan Sharma, University of Michgan

Detecting interviewer falsification is an important part of quality control in many 3MC surveys. However, real-world case-studies on this aspect are uncommon. In this presentation, I talk about a television audience measurement panel survey in India that produced viewing estimates widely used as 'currency' to transact advertising business. Over the years, there were reports of attempts being made to unethically influence respondents' viewing - via interviewers or directly - so that estimates could favor certain television channels. Detecting possible falsification therefore became a critical quality control activity, consisting of analysts looking at data to spot suspicious changes in viewing behavior. Using a real case, I show how multilevel models can largely replace such methods by offering three advantages. First, the model approximates an interpenetrated design by simultaneously controlling for several household characteristics. Second, the model allows us to examine effects at the household and interviewer levels separately so that one can direct investigation efforts accordingly. Third, by using a formal yet simple model, we obtain a more objective and faster method. We suggest that such models be part of the regular survey quality control toolkit.

Wednesday, July 27th, 2:00 p.m. - 3:30 p.m.
Session: Questionnaire Translation
Chair: Alisú Schoua-Glusberg, Research Support Services
Location: Great Lakes A/B

Trend measurement in international assessment surveys from a linguistic quality assurance perspective: To repress, to encourage or to manage the urge to improve trend items in translated instruments
Laura Wayrynen, cApStAn
Beatrice Halleux, HallStat
Andrea Ferrari, cApStAn

In international surveys, investigators collect data about a given population at a given point in time. In addition, if periodical data collections are planned, it is of great interest to also measure change over time. For this purpose, some questions are administered repeatedly across survey waves. We shall refer to these questions as trend items. Al Beaton famously said if you want to measure change, don't change the measure, and that is a starting point in studies such as PISA or TIMSS. This seems a sensible approach, since there is abundant literature about the impact of even minor changes on item statistics. However, questions can become outdated due to a spelling reform in the target language; an educational reform in the target country; or contextual change such as a change of currency in the target country. Also, errors that went undetected in the first wave may be discovered when preparing the next wave. Finally, the delivery mode may have changed, e.g. from PAPI to CAPI. Another risk factor is that project teams in participating countries are inclined to review trend items. In such cases, it can be challenging to assess whether the proposed changes are preferential edits (that may represent linguistic improvements or not); whether they correct outright errors; or whether they are necessary modifications due to a change in local context. In any case, it is necessary to design strict procedures to filter and control changes in trend content, so that even the tiniest edit is clearly documented and its effect can be tracked. In an ideal world, such requests for changes should always be supported by data: if an item had a country/item or language/item interaction that could be described as differential item functioning, there is a good reason to scrutinize the wording or cultural adaptations and, possibly, to propose alternative wording. Conversely, one could advocate that if an item worked well (i.e. the translated item has been fielded and has not shown any unusual statistics), then correcting a residual error may be an unnecessary risk. This paper examines different approaches used in projects in which the authors are involved as field practitioners, and the advantages and drawbacks of each approach will be analysed.

The Translation Management Tool (TMT) – a single online platform for translating cross-cultural survey instruments

Brita Dorer, GESIS-Leibninz-Institute for the Social Sciences
Maurice Martens, CentERdata

The Translation Management Tool (TMT) is an online service for supporting translation processes for large multilingual surveys. It has been used since 2004 for the renowned Survey of Health, Ageing and Retirement in Europe (SHARE) and over time has supported several other studies. In order to be useable also for surveys that apply the 'team' or 'committee approach' for their questionnaire translations, it is currently being adapted for use by the European Social Survey (ESS). The ESS has used the team approach , in the even more elaborate form of a 'TRAPD model' , since its very beginning. The paper will present the current state of the TMT tool and to what extent it is useable for the ESS translation process. Challenges from both authors' perspectives will be presented and discussed: from the ESS translation team's side in terms of how smoothly the tool can be integrated into the ESS processes: the different roles contributing to the ESS translations, such as translators, reviewers, project managers or external verifiers , all need to use the tool without unnecessarily delaying the whole process. The ESS workflow is designed on a very global level, that is, the whole questionnaire is always in a certain 'state' (e.g. translation, review, verification), while TMT historically defines its statuses at a lower, that is, question level. The TMT can be configured to set up translation processes dynamically and therefore adjust itself to support existing translation processes. The paper will discuss how to support multiple distinct workflows and definition levels in one tool. It is planned, that, at a later point in time, this TMT should be interlinkable with two other tools using DDI3: the 'QDDT' to document the questionnaire development process, and the 'QVDB', a question and variables database. In the end all three tools together will be useable for carrying out and documenting the whole questionnaire design and translation process, including storing final translations in a searchable database. The TMT will be the first questionnaire translation tool available that can be used for questionnaire translations carrying out the team approach; in addition, it will form the cornerstone within this series of tools useable for the whole questionnaire development and translation process.

* Advantages and limitations of documentation of a sophisticated survey translation and adaptation process

Dorothée Behr, GESIS
Steve Dept, cApStAn
Elica Krajceva, cApStAn

In Chapter 9 of the Wiley book (Dept, Ferrari, & Wäyrynen, 2010) that resulted from the 3MC Conference in 2008, a case was made for a centralised monitoring tool for questionnaire translation. The authors pleaded for a repository in which each translation/adaptation step would be documented for each country/language version of each item. Such a tool would not only contain the different versions across the entire production cycle of a questionnaire translation but also comments in case of noteworthy problems or decisions, the need for which has repeatedly been stressed (Harkness, Pennel, & Schoua-Glusberg, 2004).

EU-OSHA (European Agency for Safety and Health at Work) gave the vision of a central monitoring tool concrete expression in its Second European Survey of Enterprises on New and Emerging Risks (ESENER-2): TNS-Sozialforschung and cApStAn were commissioned to set up a translation and linguistic quality assurance design based on the TRAPD survey translation process (Harkness, 2003). The key characteristics of the ESENER-2 study were as follows: (i) a single English-language source questionnaire; (ii) an attempt to implement best practice in survey translation, which involved translatability assessment, double translation, individual reconciliation as well as team adjudication, domain expert assessment, proofreading, and piloting; (iii) documentation of each step in the process in an Excel file; (iv) input-related documentation included item-specific conceptual notes and translation instructions; (v) output-related documentation included both the actual translation and comments in case of particular problems or decisions; (vi) the process was centrally managed by cApStAn; (vii) the same process was used for 47 national language versions.

EU-OSHA and TNS authorized the authors of this presentation to regard the contents of all 47 TAFFs as data, to analyse this data, and to report on it. This paper will focus on the usefulness and the limitations of comprehensive documentation. First, a framework of translation documentation will be presented, which differentiates between input documentation that is fed into the process to support translation and output documentation that is produced throughout the process. This includes the translations themselves, any comments, queries and answers. In a next step, the ESENER-2 study will be presented, including its various steps and the documentation produced at various points in the translation process. What follows is an in-depth look at the various types of documentation and overall project set-up from the perspectives of project management, translation and translation research.

Up to now, there has been relatively little insight on the best way to collect, organise and make good use of translation documentation (Zavala-Rojas & Saris, 2014). This presentation will give a comprehensive account on what documentation involves, how it is used, and how it may be used in the future. It thus adds empirical data to the previously rather prescriptive stance on documentation and, beyond that, offers suggestions for future use and streamlining of documentation. Further, it brings together methodologists and field practitioners and thus allows theory and practice to inform each other.

Wednesday, July 27th, 2:00 p.m. - 3:30 p.m.
Session: Response Scales 1

Chair: Sunghee Lee, University of Michigan
Location: Michigan Ballroom I

*** Correcting for differential response scale usage across cultures through anchoring vignettes**

Mengyao Hu, University of Michigan
Sunghee Lee, University of Michigan
Hongwei Xu, University of Michigan

In cross-cultural studies, one major source of measurement noncomparability in self-assessments is reporting heterogeneity that arises due to differences in respondents usage of response scales typically those using ordinal quantifying descriptors (e.g., extreme-severe-moderate-mild-none). Regardless whether it is due to cultural differences or questionnaire translations, this adds complexity to interpreting the results from cross-cultural comparisons, often dampening measurement comparability and leading to measurement error. One method developed for correcting for reporting heterogeneity in cross-cultural comparisons is the anchoring vignette method. The data for this method have been collected in a wide range of surveys, including the World Health Surveys (WHS), the Study on Global Ageing and Adult Health (SAGE), the Health Retirement Survey (HRS), the Survey of Health, Ageing and Retirement in Europe (SHARE), the English Longitudinal Study of Ageing (ELSA) and the Chinese Health and Retirement Longitudinal Study (CHARLS) to name a few. This methods starts from the premise that, for a given domain, there is a true and unobservable state of each person that lies on a continuum of the domain, that in data collection, respondents use unobservable cut-points that enable them to report their state on a ordinal response scale and that respondents cultural backgrounds interact with where their unobserved cut-points are located systematically. The potential differences in cut-point locations across cultural groups result in reporting heterogeneity. This method attempts to investigate where the cut-points are located by giving respondents a few vignette items that describe hypothetical persons situations related to the domain of interest following a self-assessment question. Then respondents are asked to assess the vignette person's state. By having respondents assess the same set of vignette items and assuming they will apply the same cut-points as in self-assessment, this method allows researchers to compare where respondent's self-assessment stands relative to their assessments on vignette persons. In other words, vignettes help reveal where respondents anchoring points lie on the continuum of the true state, which, therefore, enables the correction for the differential response scale usage in cross-cultural comparisons. This paper aims to provide an overview of the anchoring vignette method and its usage in cross-cultural studies in the following three sections: 1) a review of the current anchoring vignette literature; 2) an empirical demonstration of

the anchoring vignette method using several data sources; and 3) a discussion on critical assumptions and practical considerations regarding anchoring vignette administration in survey data collection. In the first section, we will introduce the background of the anchoring vignette method and vignette items developed for various domains and how the method has been applied in the current literature and discuss the required assumptions. In the second section of this paper, we will demonstrate an empirical application of using anchoring vignette data to correct for reporting heterogeneity in several selected countries and their subgroups. For example, we will use data from aging-population surveys, such as the HRS, SHARE, ELSA, CHRLS and SAGE. These surveys include self-assessments on several health domains, including sleep, mobility, concentration, breathing and affect, along with anchoring vignette data for each domain. Group comparisons will be carried out in three stages. In the first stage, we will compare the domains of interests across selected groups solely based on simple comparisons of the self-assessments. In the second stage, we will make comparisons of self-assessments after controlling for background variables through multivariate models (e.g., probit). In the third stage, anchoring vignette data will be added to the multivariate models to correct for potential reporting heterogeneity across selected groups in making comparisons. Comparing the results from these three stages will allow us to examine whether reporting heterogeneity exists and the pattern of group comparisons before and after taking reporting heterogeneity into considerations through anchoring vignette data. The third section of this paper will discuss practical issues of implementing and administering anchoring vignette questions in survey data collection and using anchoring vignette data. We will demonstrate these issues through empirical investigations using data that we collected specifically for methodological assessments of anchoring vignettes, including survey administration time, cognitive difficulty, and order of vignette questions. This section will also include discussions on the sensitivity of critical assumptions in using anchoring vignette data. We expect this paper to provide a comprehensive overview of the anchoring vignette method to survey researchers for a better understanding: what the method is about, what it can and cannot do, how to apply it in cross-cultural studies, and most importantly, rarely discussed practical considerations and assumption tests.

Measurement equivalence of agree/disagree and item-specific response options

Natalja Menold, GESIS
Anna Andreenkova, CESSI

Item-specific response format uses categories related to the evaluation domain of a question. With the question How would you rate your health , excellent, very good, good, fair, or bad , an example for item specific response format is given. This item can alternatively be asked with agree-disagree format, which can be used with any evaluation domain: To what extent do you agree strongly or disagree strongly that your health is excellent with a rating scale ranging from strongly agree to strongly disagree. This example is taken from Saris et al. (2010). In this article the authors show that in the European Social Survey (ESS) item-specific format is associated with a higher measurement quality. In addition, it has been argued in the literature that item-specific format is less prone to acquiescence than agree-disagree format. With our analyses we test the hypotheses using ESS data that item-specific format is rather associated with measurement equivalence across countries and time than agree-disagree format. Measurement invariance is crucial when comparing statistically obtained parameters between different groups or points of time. The results will be presented and discussed with respect to the cross-cultural comparability of the ESS data.

* Response Styles in cross-cultural surveys: A tutorial on estimation and adjustment methods and empirical applications

Z. Tuba Suzer-Gurtekin, University of Michigan
Mingnan Liu, SurveyMonkey
Florian Keusch, University of Mannheim
Sunghee Lee, University of Michigan

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. Two of the most frequently studied response styles are acquiescent response style (ARS) and extreme response style (ERS). Focusing on ARS and 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.

Methodological approach for the development of equivalent rating scales for comparative cross-national surveys

Anna Andreenkova, CESSI
Natalja Menold, GESIS

In the recent decades cross-country comparisons based on survey data are widely used for academic analysis and for policy-making. To make reliable comparisons, survey data should satisfy requirements of comparability on measurement level. One of the most broadly used question design for measuring attitudes, values and opinions of general population is questions with rating scales with verbal labels as answer categories. Regardless of wide use, comparability of labeled rating scales is one of the least studied issue in comparative surveys research methodology. The lack of equivalence in rating scales and verbalization of rating scales can limit cross cultural comparability (Mohler et al., 1998), prevents from drawing reliable conclusions or can even mislead the analysis. We suggest the approach to construct equivalent rating scales for comparative surveys of different countries and languages based on rigorous methodological experimental design used for CRS-GR project (Comparable Rating Scale between German and Russian project). The approach relies of simultaneous construction of rating scales in different languages and parallel evaluation of the equivalence of these scales. First stage of designing equivalent scales is explorative analysis of available data and practice along with linguistic analysis (linguistic characteristics of each label including strength, emotional power, fixed or unfixed nature, etc). On second stage of the project, cognitive interviews are used to empirically explore the perceived meaning of each item of rating scale, subjective distance between items, major characteristics of each labeled item for respondents from different social groups, education, gender and age. Based on the analysis of cognitive interviews, few rating scales are selected for quantitative test to obtain the subjective score (distance) for each label obtained and the comparison of the most equivalent labels between different languages. As a final step, experimental test is used to evaluate several alternative verbal scales which have the potential for being highly comparable between languages to check internal consistency, concept equivalence and concept comparability. As a result of this procedure few rating scales for different topics are constructed with empirically proved equivalence between countries and languages.

Wednesday, July 27th, 2:00 p.m. - 3:30 p.m.

Session: Sampling Approaches

Chair: Ting Yan, Westat
Location: Great Lakes D

*** Within-household selection of respondents: The last step of sampling in household surveys**

Achim Koch, GESIS

In surveys of the general population, sampling often comprises a two-step process. First a sample of households has to be drawn, and second a respondent has to be selected from the eligible members in each household. This chapter provides an overview on within-household selection methods, focusing on approaches which can be classified as probability or quasi-probability methods: the Kish method and the birthday technique. The European Social Survey serves as an example to provide some empirical evidence on the use and resulting quality of different within-household selection techniques in a cross-national survey. The question of utilizing within-household selection methods in non-Western countries is discussed, and basic recommendations for the use of within-household selection techniques in cross-national surveys are provided.

Developments in sampling and efforts to achieve comparable sample estimates for the European Social Survey

Stefan Zins, GESIS
Siegfried Gabler, GESIS

The European Social Survey (ESS) has devoted a considerable effort to the task of obtaining probability samples of comparable quality. To achieve this the ESS has installed a system of sample management that starts form the planning of the sampling design, to its implementation, and finally the recording of sampling related meta-data that is used to calculated survey weights and to assess sampling related quality aspects of the gathered data. At the centre of the sample management is a group of sampling experts that can draw from the accumulated experience over all previews ESS rounds. The work reflects on this experience of over-seeing the planning and implementation of sample designs and how outcomes have improved over the years, but also some of the difficulties encountered will be discussed. An integral part of the efforts to harmonise the quality of the national samples are comparable sampling errors. To both control for the sampling error and still allow for sampling designs that suit the capabilities of the participating countries to select probability based samples and conduct face-to-face interviews, the concept of a minimal effective sample sizes is used, which every country has to achieve with its sample. This includes the preparation of benchmarks values of country specific design effects, which are essential to the planning of the sample sizes. With its seven rounds the ESS has for some countries a rich history of data and meta-data on sampling that is used to map the evolution of design effects and its driving factors. This analysis helps to identify and to target certain aspects of the sampling design for actions to further improve the accuracy of estimates based on ESS data. The development of sampling designs, regarding sampling techniques used and available sampling frames, and how this has affected the design effects will be made apparent to highlight actions on improving the quality of the ESS data. The goal is to facilitate the ongoing efforts in the ESS to assure the achievement of the minimal required effective sample sizes while at the same time keeping the costs of the survey checked.

Design effects from a geographic sampling-based study of barriers to internet access in developing countries

Jennifer Unangst, RTI International
Jeniffer Iriondo-Perez, RTI International
Safaa Amer, RTI International

A multinational study was conducted across various developing countries in Africa, Asia, and South America to identify barriers to Internet access across gender and socio-economic levels. This study was based on a complex probability-based geographic sampling (geo-sampling) design which was standardized across countries as much as possible and representative at the country level. However, country-specific tailoring was needed due to differences in administrative structure across countries and inconsistency of data available on the distribution of the population to support the design. Examples of modifications to the standard design included stratification of the primary sampling units by urban/rural or poverty level, varying the number of sampling stages, allocation of the sample at each stage of the design, and sampling strategy (e.g. probability proportional to size, systematic sampling, etc.). This paper discusses the weighting approach to account for the complex design and associated challenges across countries for the geo-sampling design, as well as the impact of country-specific adjustments on the sampling error and design

effect. In addition, the paper uses simulation study results to assess the impact of some simplifying assumptions made for weight calculations on the final weights and design effect. This is specifically linked to a step in the sampling strategy which assumes a uniform distribution of the population across areas in rural settings and excludes non-residential sampling units sequentially at the stage of selecting the smallest sampling area. The simulation studies the weight calculation at this stage by comparing binomial distribution to geometric distribution assumptions. Finally, the paper describes the sensitivity of the results to different assumptions across some of the study's main indicators such as access to internet, usage, and satisfaction measures.

* Sample designs Using GIS technology for household surveys in the developing world

Stephanie Eckman, RTI International
Kristen Himelein, World Bank
Jill Dever, RTI International

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.

Case of dual frames in indirect sampling

Manuela Maia, Universidade Católica Portuguesa

Multiple frame design is a strategy that deals with the problem of under coverage of sampling frames, which consists in combining several frames in order to provide complete or nearly complete coverage of the target population. In most cases, the frames overlap, causing a problem to estimate in what regards sample weights computation. Therefore, Indirect sampling can be an alternative approach to the classical sampling theory in dealing with the overlapping problem of sampling frames on survey estimates.

In this paper, the classical estimators of multiple frames sampling - Domain Membership estimator and Unit Multiplicity estimator – are translated to the context of indirect sampling. Additionally the Optimal Deville and Lavallée estimator is decoded to the context of multiple frames surveys. The purpose is to deduce a new class of indirect sampling estimators capable of being applied in multiple frames surveys, more specifically in the particular case of dual frame surveys.

The new estimators are then compared with the indirect sampling of Optimal Deville and Lavallée estimator, under eight different scenarios of links between sampling frame and target population, in order to identify which of them is more efficient. Henceforth, both theoretical comparisons and comparisons using simulation reveal that the Unit Multiplicity estimator and the Optimal Deville and Lavallée estimator are equally competent and are more efficient than the Domain Membership estimator.

Wednesday, July 27th, 4:00 p.m. - 5:30 p.m.
Session: Cognitive Interviewing

Chair: Alisú Schoua-Glusberg, Research Support Services
Location: Great Lakes A/B

Approaches to integrate cognitive interviewing data and psychometrics for assessing bias from linguistic, context and culture factors in 3MC survey research

Jose-Luis Padilla, University of Granada, Spain
Isabel Benitez, Loyola University Andalusia, Spain
Fons van de Vijver, Tilburg University, The Netherlands. North-West University, South Africa. University of Queensland, Australia

Cognitive interviewing combined with psychometrics in a mixed methods study is a powerful strategy to examine question/item bias in cross-cultural research. There are different approaches to integrate quantitative and qualitative data. The aim of this paper is to examine how CI data and psychometrics can be integrated at different levels: design, methods, interpretation and reporting, to detect and interpret bias in cross-cultural research. We illustrate our proposals by presenting a CI study carried out within a mixed methods design aimed at assessing bias in Quality-of-Life (QoL) scales, used in large-scale quantitative studies. A total of 50 participants, 25 participants from Spain and 25 from the Netherlands responded to QoL scales examined and then took part in the cognitive interviewing study.Demographic profiles of CI participants were set to match characteristic of samples in the quantitative study. Experienced interviewers conducted interviews in mother language's participants following interview protocols in Spanish and Dutch. Subthemes and themes were developed through the iterative analytic process. Saturation was reached before all interviews were finished. The integration of CI findings and quantitative results suggested three main sources of bias in QoL measures for Spanish and Dutch respondents: linguistic issues, content of questions, and contextual factors. CI findings were informative to understand psychometrics of QoL item scales. Lessons learned while integrating CI data and psychometrics within a mixed-method study will be also exposed, as well as the benefits of the CI analytic techniques implemented for improving comparability in cross-cultural survey research.

The practice of cross-cultural cognitive interviewing

Gordon Willis, National Cancer Institute, National Institutes of Health

In a recent review article entitled The Practice of Cross-Cultural Cognitive Interviewing: A Research Synthesis, Willis (2015) summarized the state of the science in this key area of comparative survey research (Public Opinion Quarterly 2015 79 (S1): 359-395). For the proposed presentation, I will summarize the major themes of this article in order to promote discussion of these ideas; to elicit 3MC participant descriptions of further related work that they may have done in application to Multinational, Multiregional and Multicultural contexts; and to promote the conduct of further research in the area of Cross-Cultural Cognitive Interviewing. The abstract of the POQ paper, which will form the basis of the presentation, is as follows: Cross-cultural cognitive interviewing (CCCI) has increasingly been practiced across a range of cultures, languages, and countries, in an effort to establish cross-cultural equivalence of survey questions and other materials, to detect sources of difficulties in answering survey questions for particular subgroups, and to detect problems related to translation from source to target languages. Although descriptions of such studies have proliferated in both the published and unpublished literatures, there has been little effort to reconcile discrepant views, approaches, and findings. The current synthesis reviews 32 CCCI studies located in peer-reviewed journals and books, along with key unpublished sources, to characterize these investigations in terms of their purpose, procedures, and findings. Based on a number of trends in this emergent field, conclusions are made concerning appropriate methods for cognitive testing of cross-cultural instruments, and recommendations are made for future practices that will serve to advance the CCCI field.

Putting yourself in respondents' shoes: Cognitive testing for cross-cultural interviewing
Kyle Block, D3 Systems
Stacey Frank, D3 Systems

Have you ever reviewed survey data from an evaluation in a foreign country and wondered how, or even if, the respondent understood the question? Many international evaluators underestimate the cognitive complexity of responding to survey questions. This session will help you break it down, describing the four primary cognitive steps involved in answering questions: comprehension, memory retrieval, judgment and estimation, and reporting. Armed with a fuller understanding of these cognitive steps, evaluators will be equipped to avoid common cognitive breakdowns and write survey questions with less measurement error. The presenters will draw on their years as international evaluation practitioners to highlight common missteps and provide solutions for writing questions that will be cognitively equivalent in a variety of cultures. This session will also introduce cognitive interviewing as a technique for uncovering misunderstandings and cross-cultural misinterpretations. Participants will have the chance to practice this skill by conducting a mock cognitive interview. This session will introduce participants to Cognitive Aspects of Survey Methodology (CASM) research, which seeks to connect the fields of cognitive psychology and survey research. These two fields converge in the questionnaire design phase of an evaluation. An evaluator must understand the cognitive process that a respondent goes through when answering survey questions so she can design questionnaires that reduce measurement error and accurately collect the data needed to answer the evaluation questions.

English and Spanish testing of questions for the National Health and Nutrition Examination Survey (NHANES): Results and lessons learned
Meredith Massey, National Center for Health Statistics

Despite the fact that many U.S. surveys are fielded in both English and Spanish, cognitive pre-testing of these instruments is not always conducted in both languages. While pre-testing in English may be emerging as standard practice, pre-testing in Spanish poses a different set of challenges in terms of recruitment, interviewing and analysis. This presentation discusses the lessons learned from a project undertaken by the Collaborating Center for Questionnaire Design and Evaluation Research at the National Center for Health Statistics (CCQDER/NCHS) to cognitively test both the English and Spanish versions of proposed new questions for the National Health and Nutrition Examination Survey (NHANES). The questions were related to consumer behavior in the context of nutrition information posted on menus and consumer product labels. Cognitive interviews were conducted with a total of 55 individuals: 40 in English and 15 in Spanish. A purposive sample was recruited of adults who met screening criteria relevant to the question topics. Detailed records were kept on recruitment strategies and rates, notes were compiled on interview debriefings, and analysts used Q-Notes, an analysis software tool, to facilitate data organization and analysis. The presentation will draw upon findings from these recruitment records, interviewer debriefings and analytical findings. Results and discussion will focus on lessons learned, recommendations, and implications for recruiting Spanish-speaking respondents with limited English proficiency; administering cognitive interview probes in Spanish; and the logistics of cognitive interview projects conducted in multiple languages. Additionally, results will be presented demonstrating how testing in Spanish can influence improvements to both the English and Spanish language versions of the instrument.

Wednesday, July 27th, 4:00 p.m. - 5:30 p.m.
Session: Measurement Invariance
Chair: Jose-Luis Padilla, University of Granada
Location: Great Lakes D

Explaining cross-national inequivalence in factor loadings and intercepts: A Bayesian multilevel SEM approach

Bart Meuleman, University of Leuven

In comparative research, the importance measurement equivalence is increasingly acknowledged, and various statistical toolkits (multigroup CFA, latent class analysis, IRT models) are readily available for applied researchers. There exists less agreement, however, on how to proceed when equivalence is found to be absent - as is often the case in practice. A highly interesting suggestion is Poortinga's (1989) proposal to treat inequivalence as a useful piece of information on cross-cultural differences. Based on this idea, Davidov et al. (2012) have introduced a multilevel structural equation modelling (MLSEM) approach that can be used to interpret deviations from scalar equivalence substantively by modelling how cross-national differences in item intercepts are linked to contextual variables. This paper travels further along that road, by extending the work of Davidov et al. (2012) in two respects. First, while previous work only considers scalar inequivalence (i.e. cross-country differences in item intercepts), this paper proposes a model that also incorporates metric inequivalence (i.e. cross-country differences in factor loadings). Furthermore, we apply a novel Bayesian framework for MLSEM that is better suited to deal with the specific features of cross-national data, to wit a relatively small sample size at the country-level. Data from the 'national identity' module of the International Social Survey Program (ISSP) 2013 is used to illustrate the models.

Cross-cultural measurement invariance among German migrants in welfare benefits receipt

Jonas Beste, Institue of Employment Research
Arne Bethmann, Institue of Employment Research

An emphasis of many surveys is the measuring of subjective indicators concerning a wide field of topics. The measurement instruments used for these purposes (e. g. batteries of multiple items) rely on the assumption of measurement invariance. This means, that all respondents have a similar understanding of the measured underlying construct as well as each individual item. To compare means of different groups of respondents we must ensure that these groups understand and respond to the questions in similar ways. Otherwise comparison between groups can lead to incorrect conclusions. Previous methodological research has shown that measurement invariance is not given for all instruments and groups. Particularly, differences appear between groups of different cultural background (Davidov et al., 2014). Therefore testing measuring invariance is of utmost importance for surveys including respondents with cultural diversity. The Panel Study Labour Market and Social Security (PASS) is an ongoing yearly household panel study of German welfare benefits recipients and is concerned with their living conditions, socio-economic situation and the dynamics of welfare receipt. Culturally the PASS respondents are very heterogeneous due to the large proportion of individuals with a migration background. This is intensified by the rising number of refugees from Syria, Iraq and Afghanistan over the last month, which enter the study through annual refreshment samples. The recent developments and their socio-political implications increase the need for valid sociological insights. To assure comparability between groups of different cultural background we test multiple measurement instruments for multiple subjective indicators in PASS (e. g. well-being, life satisfaction, social participation and inclusion, psychometric measures) using a multi-group CFA framework (see Vandenberg & Lance 2000). We operationalize cultural background using questions on migration background, spoken language, religion and nationality as well as the language the interview was conducted in.

Cross- and within country measurement invariance in a EU comparative research on school dropout

Ward Nouwen, University of Antwerp (Belgium)

As part of the five year Reducing Early School Leaving in Europe comparative research project (www.resl-eu.org), a longitudinal student survey was developed and administered across fourteen urban areas in seven EU member states (i.e. Belgium, Spain, The Netherlands, Poland, Portugal, Sweden and the United Kingdom). In total 19631 students in urban secondary schools were questioned in the first wave of the survey during the spring of 2014. The student survey was developed based on a theoretical and methodological framework encompassing state-of-the-art

international literature on explaining school dropout. The questionnaire builds on reliability and validity testing of measurements in previous studies includes background information as well as a wide range of measurement constructs for social support and school-related attitudes and behaviour. These constructs were adjusted to the different national and linguistic contexts based on a methodological process that included translation/ back-translation, national piloting and reliability testing. Theoretically, the paper largely builds on the school engagement concept, a prominent concept in international literature on explaining school dropout. Although there is ongoing discussion on the different dimensions of the school engagement concept, a constant across the conceptualisations of school engagement is that it is multidimensional concept that contains emotional, cognitive and behavioural components. In the multidimensional operationalisation of the concept proposed by Fredricks, Blumenfeld & Paris (2004) behavioural engagement consists of the actions and practices students direct toward school and learning , including both positive and negative behaviour. The emotional component represents a student's sense of belonging to school and valuing of education. Cognitive engagement refers to a student's self-regulated and strategic approach to learning. For the operationalisation of Fredericks et al. (2014) multidimensional operationalisation of school engagement, the RESL.eu project builds on Wang, Willett & Eccles' (2011) multidimensional assessment of school engagement. Methodologically, the paper will present measurement invariance testing using structural equation modelling and focusses on an adapted assessment of the multidimensional school engagement across the urban student populations in seven EU member states. Furthermore, the paper will compare the level of measurement invariance between and within the same national educational systems by also performing measurement invariance testing between different educational tracks for those partner countries that provide non-comprehensive upper secondary education. Next to cross-country differences, measurement invariance testing in the socially stratified tracked Flemish (Belgian) educational system showed that measurement invariance issues across different educational tracks within the same country is an issue that cannot be neglected in multi-cultural survey methodology.

Wednesday, July 27th, 4:00 p.m. - 5:30 p.m.
Session: Mixed Mode/Methods

Chair: Michèle Ernst Stähli, FORS
Location: Michigan Ballroom II

*** How to design and implement mixed-mode surveys in cross-national surveys: Overview and guidelines**

Edith D. de Leeuw, Utrecht University
Ana Villar, City University London
Z. Tuba Suzer-Gurtekin, University of Michigan

The single mode paradigm, which implies that one data collection mode fits all respondents perfectly, is no longer tenable for all survey research purposes (de Leeuw, 2005). Within countries, surveys increasingly use mixed mode data collection (e.g., combining online surveys and interviews) because this can help to control costs and to maintain good response rates (Biemer & Lyberg, 2003). For instance, a serious threat for web-mode only is undercoverage (Bethlehem & Biffignandi, 2012). Penetration of new technologies (e.g., internet connections, smartphone, i-Pad or tablet), differs between subgroups, the so called digital divide (Couper, 2000, 2008; Mohorko et al, 2013a & b). A combination of different survey modes in one study, be it cross-sectional or longitudinal, can reduce undercoverage of specific subgroups (e.g., elderly or lower educated). Mixed-mode surveys appear to be appealing in reducing important survey errors at reasonably costs. However, a combination of different survey modes in one study may lead to different kinds of measurement errors (Hox, De Leeuw, & Dillman, 2008; De Leeuw & Hox, 2011; Tourangeau, Conrad, & Couper, 2013). In an international setting, mixed mode studies are sometimes inevitable, and in the words of Blyth (2008) the only fittness regime. Countries differ in overall telephone coverage and in the amount of cell phone only households (Mohorko et al, 2013a); in Internet penetration rates (Behlehem & Biffignandi, 2012; Mohorko et al, 2013b), and in social and economic circumstances, which influences the general economic situation of National Statistical Institutes and survey organizations, and may limit the data collection modes used within a country (Blanke & Luiten, 2014). Finally countries differ in survey tradition, customs, and in literacy rates (Skjak & Harkness,

2003; Harkness et al, 2010; Pennell et al, 2010). These differences can lead to different countries implementing different data collection designs to better tailor the country's situation, and such mode differences may threaten the cross-cultural comparability of data. Modes may differ in many aspects (Roberts, 2007; Jackle et al, 2010). Still there are only a few mode-inherent factors on which survey modes differ (Berzelak, 2014), such a closeness of interaction with interviewer and use of computer technology; however many mode differences are not inherent, but design specific. Different data collection designs have different consequences on a survey's coverage error, nonresponse error, and measurement error (Tourangeau, 2013). As a consequence, some of the mode differences are implementation specific and as such may be reduced in the design phase (e.g., Dillman & Christian, 2005; De Leeuw et al, 2008). In an attempt to do so, the U.S. Census Bureau and the U.K. Office for National Statistics prepared guidelines for designing and implementing mixed-mode surveys for government surveys (Betts & Lound, 2010; Martin et al., 2007). In this paper we give an overview of current best methods in designing and implementing mixed-mode surveys from a cross-national perspective. We will also review and discuss the implications that mixing modes of data collection in cross-national surveys have on comparability and total survey error, using examples from studies that have compared single mode designs to mixed mode designs in cross-cultural contexts. For example, the European Social Survey has conducted studies testing the effects of telephone surveys compared to face-to-face surveys as well as mixed-mode surveys compared to face-to-face surveys. We will focus on the prevention of mode effects, whereas statistical adjustment methods fall outside the scope of this paper. Special attention will be paid to the development and implementation of questionnaires in a mixed-mode, cross-national setting, discussing trade-offs between cost savings and survey error.

* Mixed methods data collection in a comparative international context: New technologies, new opportunities for social science

Nathalie E. Williams, University of Washington

This chapter will address mixed methods data collection opportunities, with a focus on new technologies, and the variety of applications that mixing these data collection strategies opens up. The purpose here will be to realign ideas of what are mixed methods, from a myopic focus on quantitative-qualitative to include a multitude of other types of informative data that do not neatly fit this dichotomy. As this volume focuses on survey methods, this chapter will particularly focus on new types of data that can be combined with survey data, to create opportunities to examine old questions from new perspectives and also to investigate entirely new questions in the social sciences. The chapter will give concrete examples of different types of data that can recently emerging with new technologies and social science projects that have used these data. It will discuss different ways these data can be used, how well they can be used in the comparative international context, and some apparent limitations.

* Mixed-mode surveys: An overview of design, estimation and adjustment methods and empirical applications

Z. Tuba Suzer-Gurtekin, University of Michigan
Richard Valliant, University of Michigan
Steven G. Heeringa, University of Michigan
Edith D. de Leeuw, Utrecht University

Mixed mode surveys (e.g. combining web and interview-modes) are increasingly used both in cross-sectional and longitudinal studies within one country to control costs and to maintain good response rates. However, a combination of different survey modes in one study may lead to different kinds of measurement errors (Hox, De Leeuw, & Dillman, 2008; De Leeuw & Hox, 2011; Tourangeau, 2013). In cross-national research, a mixed-mode approach is sometimes inevitable (Blyth 2008), as countries differ greatly in internet coverage, telephone penetration (Mohorko et al, 2013), literacy, or survey traditions. In addition, mixed-mode surveys can be used as part of the modern enhancement strategies of response rates, yet they may cause a challenge for the comparability of data. This paper reviews three related issues in mixed mode surveys: design, estimation, and adjustment. We start with a concise overview of design issues in mixed-mode surveys from a cross-national/cross-cultural context (e.g., Harkness

et al, 2010), and discuss the importance of design and auxiliary data for estimation and adjustment (Groves & Lyberg, 2010) The emphasis of this paper is on estimation and adjustment and the evaluation of mixed-mode surveys when modes are non-randomly assigned as is the case in cross-national surveys (e.g. ISSP). In practice, survey modes are not randomly assigned in mixed-mode surveys. This nonrandom assignment (called here mode choice) establishes a challenge to evaluate different sources of variability when studying mode effects in mixed-mode surveys. Observed differences between respondents participating through different modes can be due to either selection effects (i.e., differential likelihood to participate depending on mode) or to mode effects related to measurement characteristics of the instrument (visual vs. aural channel, presence vs. absence of another person, etc). Recently, some statistical methods have been proposed to quantify and isolate mode effects from mode choice, to validate mixed-mode survey estimation assumptions (Camillo & D'Attoma, 2011; De Leeuw, 2005; Jackle, Roberts, & Lynn, 2010; Klausch, Hox, & Schouten, 2013; Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011; Suzer-Gurtekin, 2013; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010, 2012; Voogt & Saris, 2005). These methods challenge the general notion of ignorable mode effects in the mixed-mode surveys and motivate a more systematic approach to evaluate mode effects. In parallel, research emphasizes the possible differential measurement error across modes (Aichholzer, 2013; Revilla, 2010; Weijters, Schillewaert, & Geuens, 2008; Ye, Fulton, & Tourangeau, 2011). Buelens and Van den Brakel (2011) also propose a mode calibrated method for estimating changes in means over time. Four current methods are reviewed that focus on disentangling the nonrandom selection of modes, mode choice, and mode effects in mixed-mode surveys. 1) The nonrandom selection is controlled analytically in regression model methods (JÃ¤ckle et al., 2010). 2) The mixture distribution method, introduced by Vannieuwenhuyze et al. (2010, 2012), defines responses as having a mixture distribution given the mixed-mode survey design and computes the selection and mode effects for distribution parameter estimates, such as mean, of a variable of interest. Mathematically it is possible to show that parameters related to the selection and the mode effects can be quantified for a two-mode survey using a parallel single mode survey and assuming the same population and measurement properties. To do these comparisons, the mixture mode distribution method relies on two key assumptions which Vannieuwenhuyze et al. (2010) term completeness and representativity. These assumptions imply that parameter estimates obtained from the single-mode survey and the mixed-mode survey are unbiased estimates for the same survey population with respect to the non-observational survey error. In practice, this could be a strong assumption as the mixed-mode surveys are usually conducted to enhance responses under a certain budget constraint. On the other hand, the method conceptualizes the confounding nature of mode choice and mode effects. Additionally, the method enables the computation of required sample sizes to detect mode effects. 3) Propensity score matching methods (Lee & Valliant, 2007; Rosenbaum & Rubin, 1983, 1984) unconfound the mode choice and the mode effects based on the propensity score matching strata that have been formed using the available covariates. This method defines the mode effects as the mean differences between the matched groups. 4) Imputation methods (Klausch, Hox, & Schouten, 2013; Suzer-Gurtekin, 2013) propose to use imputation to detect and adjust significant differential measurement error. All these methods focus on the comparability of the survey data as opposed to determining the best performing mode. Also in a recent paper, Vannieuwenhuyze, Loosveldt, and Molenberghs (2014) classify these methods as covariate adjustment models and outline the assumptions related to the covariates in the models. Although the principles of these methods are in parallel with the requirement of data comparability in the cross-cultural research, there is no comprehensive empirical evaluation of these methods in the cross-cultural surveys. In addition to a comprehensive overview of these methods in the context of cross-cultural surveys such as European Social Survey, this paper will provide empirical examples for detecting and adjusting for mode effects.

Wednesday, July 27th, 4:00 p.m. - 5:30 p.m.

Session: Response Scales 2

Chair: Z. Tuba Suzer-Gurtekin, University of Michigan
Location: Michigan Ballroom I

Cross-national comparison on quality of response scales

Pei-shan Liao, Center for Survey Research, RCHSS, Academia Sinica

Willem E. Saris, RECSM, Universitat Pompeu Fabra
Diana Zavala-Rojas, RECSM, Universitat Pompeu Fabra
Quality of a survey measure has been an important issue, as it is closely associated with the reliability and validity of survey questions. The recently developed split-ballot multitrait-multimethod (SB-MTMM) approach has been used to evaluate the quality of questions in survey research. It aims to reduce the response burden of the classic MTMM by means of using different combinations of two methods in multiple groups. The SB-MTMM approach has been applied to the European Social Survey (ESS) to examine the quality of questions across countries, including the differences in response design and measurement errors. Information about the quality of more than 2,700 questions from different European countries and the US can be found in the program SQP2.1. Despite the widely application and research in the European countries, whether the same quality can be achieved in a different culture like Taiwan is yet unknown. Previous studies have indicated that respondents in East Asian countries, for example, tend to choose more frequently responses in the middle of the scale because of the influence of collectivism. Also within European countries and languages, differences in quality have been found. This study aims to compare the data quality, by means of reliability and validity, of different response scales using the SB-MTMM approach. By using the same questions as in the ESS, a cross-cultural comparison will be made to understand whether the studied response scales perform equally well in Taiwan compared to those in European countries. Data are drawn from the 2015 Taiwan Social Change Survey, which is a collaborator of the International Social Survey Programme (ISSP). A two-group design is adopted and the questions used for the experiment is Satisfaction from Round 1 of the ESS. Taiwanese data are collected using computer-assisted personal interview (CAPI), and the same estimation procedures for reliability, validity, and method effects will be used. Concluding remark and discussion will be provided.

Exploring drivers of acquiescent responding among ethnically diverse Latino telephone survey respondents

Rachel E. Davis, University of South Carolina
Sunghee Lee, University of Michigan
Timothy P. Johnson, University of Illinois at Chicago
Ligia I. Reyes, University of South Carolina
Chris D. Werner, University of South Carolina
Jim F. Thrasher, University of South Carolina
Ken Resnicow, University of Michigan
Frederick G. Conrad, University of Michigan
Karen E. Peterson, University of Michigan

Background: Acquiescent response style (ARS) occurs when survey respondents systematically agree with items with Likert-style response scales. ARS represents a significant threat to cross-cultural survey research, as research indicates that use of ARS differs across cultural groups. In the U.S., ARS may be of most concern when surveying Latinos. Latinos are demographically and culturally diverse, yet almost nothing is known about drivers of ARS among Latinos or Latino ethnic sub-groups. ARS is generally assumed to be unrelated to item content; however, data from our group suggest that this assumption may be invalid. This presentation will describe findings from two telephone surveys designed to better understand drivers of ARS among Latino telephone survey respondents. Methods: The first survey was conducted with 120 adults, stratified by Latino ethnicity (Mexican American/Puerto Rican/Cuban American), language (Spanish/English), and education (high school or less/more than high school). Participants were randomized to experimental conditions testing three potential influences on ARS: the number of response options (5 points/7 points/10 points); primacy vs. recency effects; and verbally offering vs. not offering but allowing a don't know response option. The second survey, which will be completed in mid-January 2016, is being conducted with 400 adults, stratified by ethnicity (Caucasian/Mexican American/Puerto Rican/Cuban American). The second survey explores whether or not ARS is related to item content and, if so, tests whether or not ARS is more likely to occur when: items trigger low-to-moderate social desirability and the desirable response direction is clear; items trigger moderate-to-high social desirability but the desirable response direction is unclear; respondents invest low effort into responding; respondents encounter item comprehension problems; respondents have no opinion about the queried

concept; and items contain negative wording. Results: Data from the first survey indicate that ARS is more prevalent for 5-point response scales, although this more limited number of response options also makes this outcome more likely. ARS was more prevalent when response options ended with strongly agree than with strongly disagree, indicating support for recency effects on ARS. Offering don't know increased the frequency that this response was selected; however, contrary to expectations, ARS was more prevalent when a don't know response was offered. Results from the second survey, which is currently in the field, will also be presented. Conclusion: Findings from this study will yield a deeper understanding of the patterns and drivers of ARS among Latino telephone survey participants.

Response option order effects in cross-cultural context: An experimental investigation with smartphones

Yongwei Yang, Google, Inc.
Mario Callegaro, Google, Inc.

The order of response options may affect how people answer rating scale questions. Response option order effect is present when changing the order of response options of a rating scale leads to differences in the distribution or functioning of individual or group of questions. Theoretical interpretations, notably satisficing, memory bias and anchor-and-adjustment have been used to explain and predict this effect under different conditions. Recent advance in visual design principles with respect to interpretive heuristics (esp. left and top mean first and up means good) adds more insights on how positioning of response options may affect answers. A number of studies have investigated the direction and size of response option order effect, but present a complex picture and all but a very few were conducted in mono-cultural fashion. However, the presence and extent of response option order effect may be affected by cultural factors in a few ways. First, interpretive heuristics, such as left means first may work differently due to varying reading conventions (e.g., right-to-left in Arabic or Hebrew). Furthermore, people within cultures where there are multiple primary languages (e.g., Hebrew and English) and multiple reading conventions (e.g., in Japan where texts may be read left-to-right-top-down horizontally or top-down-right-to-left vertically) may respond to positioning heuristics differentially. Respondents from different countries may have varying degree of exposure and familiarity to a specific type of visual design. With an increasing number of respondents taking surveys on smartphones, rating scales have to be presented vertically, which means past findings based only on horizontal displays may not generalize to this situation. It is also conceivable that smartphone users may be less susceptible to the impact of reading conventions because they have been exposed to an abundance of web contents, which are presented overwhelmingly left-to-right regardless of language. Through a series of experiments on smartphones, we investigate rating scale response option order effect across countries with different reading conventions and industry norms for answer scale designs. Following existing literature, we also consider the effect of vertical vs. horizontal display, survey topic area, number of scale points, age, gender, and education level. We incorporate a range of analytical approaches: distributional comparisons, latent structure modeling, and analysis of response latency. Based on our findings and previous literature, we discuss best practices in presenting ratings scales on smartphones, as well as for comparative analysis involving data obtained with different rating scale response option orders.

Comparing scale direction effect among Hispanics vs non-Hispanic respondents

Ting Yan, Westat
Mengyao Hu, University of Michigan

Survey literature demonstrates that the direction of a response scale can affect survey answers by drawing more answers to the beginning of the response scale. This paper attempts to carry the research on scale direction effect one step further by examining scale direction effect for Hispanic respondents and contrasting the presence and magnitude of scale direction effect to that for Non-Hispanic respondents. Data for this paper are from the 2012 American National Election Study cross-sectional component. An experiment is conducted on a set of five items measuring electoral integrity. Half of the respondents are provided a scale running from not at all often to very often and the other half received a scale running from very often to not at all often. We propose to fit a confirmatory factor analysis on these five items and to examine group equivalence across scale direction through Multiple Group

Analysis. We will then compare equivalence results for Hispanics respondents and non-Hispanic respondents. The key research question to be answered in this paper is whether scale direction affects Hispanic and non-Hispanic respondents in the same way or not.

Physical activity in three countries: How self-reports differ from measurements
Arie Kapteyn, Center for Economic and Social Research, USC

Physical activity is a prime component of health behaviour, and accurate measurement is necessary for a better understanding of what drives differences in physical activity and how this can be influenced by policy. Obtaining internationally comparable objective measures of physical exercise has proven to be very difficult given its subjective nature. We have collected new data in the UK, the Netherlands and the US, comprising both self-reports of physical activity and measurements obtained from accelerometers that respondents wore for a week. We find systematic differences between the self-reports and objective measures across different socio-economic and demographic groups as well as across countries. Compared to the two European countries, Americans seem to be more optimistic about the level of their physical activity when compared to the objective measurements.

Wednesday, July 27th, 4:00 p.m. - 5:30 p.m.
Session: Sample Management Systems in Majority Countries: Survey Management System Challenges

Chair: Brad Edwards, Westat
Location: Great Lakes E

Collecting rich paradata to monitor data collection quality
Beth-Ellen Pennell, University of Michigan
Gina-Qian Cheung, University of Michigan

This presentation will describe innovative uses of rich paradata to monitor production and quality indicators in household surveys in developing and transitional countries. Technology is increasingly facilitating new approaches to more efficient production and better quality monitoring through the collection and monitoring of rich paradata (process data). This diffusion of technology not only allows for immediate access to the survey and process data (including call records) to monitor field work quality but has also facilitated the use of other applications. These applications include self-administered modes (e.g., audio computer-assisted self-interview [ACASI]), digital audio recordings, global positioning systems (GPS) for collecting contextual information or live tracking of interviewer travel to households, areal photography for sample selection, and the collection of various anthropometric data using digital devices, among other examples. With these innovations come new challenges, however. The presentation will discuss advantages, challenges and lessons learned across a diverse set of projects as well as make recommendations for new developments in this space. The presentation will provide examples from projects in four very different settings: China, Ghana, India, and Nepal.

Mobile data collection and reporting across multiple developing African countries
Rick Mitchell, Westat
Abie Reifer, Westat

Westat is utilizing mobile technology in support of data collection for multiple large survey projects in developing African countries, This technology has led to improvements in study activities in data collection for household listings and interviewing for surveys in the research areas of HIV, nutrition, and food security. Routine transfer of data from remote areas in the field has allowed for daily dashboard reporting of project study statuses and the sharing of key results to US partners, US government agencies, and multiple in-country African partners. The use of open-source tools such as Open Data (ODK) has allowed for stronger team collaborations with fewer obstacles in capacity building

efforts. ODK consists of client software that runs on Android devices, server software that pushes instrument definitions to the field devices and receives and stores the collected data from these devices, and tools used to author the surveys. These data collection efforts have been implemented on over 2,000 Android tablets where configurations have involved not only US partners but most importantly those in-country partners wanting to learn, grow, and have the opportunity to build on these experiences in future projects down the road. In this presentation, we will describe our experiences with multiple international data collection project efforts including both capabilities and limitations, how ODK was configured and used by Westat, data flows into and out of the ODK environment, and some of the issues to consider in adopting ODK to support survey operations.

TBD
Michael Wild, World Bank

Survey at the crossroads: Implementing electronic survey management on a large infrastructure survey in Kenya
Sarah Hughes, Mathematica

Survey planners working in developing countries in the mid-2010s face the dilemma of using a familiar paper-based survey management system or introducing an unfamiliar electronic survey management system. While most survey directors recognize the advantages of electronic sample management and data capture, such as; lower error rates; improved interviewer oversight; more flexibility in responding to changing field conditions; and a shorter time lag from data collection to analysis than when using paper, obstacles to adoption of electronic data capture persist. These obstacles include; concerns by funders and local teams about cost of hardware and software; lack of human resources for programming and maintaining data systems; changes in roles and responsibilities of field and central office staff; longer and more costly fieldwork training; and security of personnel and equipment.
The Kenya State of the Cities Baseline Survey, sponsored by the World Bank in 2012-2013, collected household listing data from approximately 153,000 households and included interviews with 14,600 households using a web-based survey management system with offline data collection capability. This presentation begins with a brief description of the survey setting, including a description of the local data collection team's capacity and experience, and an overview of the implementation of the full household listing, case selection, field management and quality oversight processes. The presentation then focuses on the challenges of implementing a very large-scale multi-language survey that introduces new systems, procedures, equipment and expectations in a resource-limited environment. The presentation will include a discussion of survey methodology capacity-building, funder expectations and constraints regarding adoption of new technologies, and lessons for survey planners adopting electronic survey management systems in developing countries.

Thursday, July 28th, 9:00 a.m. - 10:30 a.m.
Session: Analysis Methods and Tools 1
Chair: Timothy P. Johnson, University of Illinois at Chicago
Location: Michigan Ballroom II

Design effect in small-scale anthropometric surveys: What parameters influence design effect, and how does this impact survey design?
Erin N Hulland, CDC
Oleg O Bilukha, CDC
Curtis J Blanton, CDC
Eva Z Leidman, CDC

Introduction: Nutritional status of the affected population, particularly children, is a key benchmark of severity in an emergency. Small-scale cross-sectional surveys are used to provide rapid but representative estimates of nutrition

indicators. Given the gains in cost and efficiency, cluster sampling is common. For cluster surveys, an accurate estimate of the design effect (DEFF) is critical to calculate a sample size that achieves adequate precision with the minimum number of sampling units. However, despite its routine calculation, there is little empirical research into the factors that influence DEFF. Methods: We compiled data from cross-sectional cluster surveys that included an assessment of anthropometry for children 6 to 59 months conducted between 2006 and 2013, excluding surveys with fewer than 25 clusters and sample sizes smaller than 150 or exceeding 1500. DEFF for weight-for-height z-scores (WHZ) was an outcome of interest. Prevalence of global acute malnutrition (GAM) based on WHZ, mean WHZ, standard deviation of WHZ, mean and variance of the cluster size, and number of clusters were calculated for each survey; survey year and region were also recorded. A generalized linear regression model was run using backward selection to identify variables associated with DEFF. Results: 378 surveys from 32 countries were included in the analysis. WHZ DEFF ranged from 1.0 to 5.2 with a median of 1.4 (IQR= 1.1-1.7). DEFF for more than 85% of surveys was less than 2.0, yielding a right-skewed distribution. DEFF increased significantly with increase in standard deviation of WHZs (p=0.014), prevalence of GAM.

Institutional trust and crisis: A comparative analysis across time

Lluis Coromina i Soler, University of Girona
Edurne Bartolome Peral, University of Deusto

Political trust is considered as a key factor related with social capital and democracy in any society (Offe 1995, Warren 1999, Hardin 1999, Uslaner 2009). The extent to which people trust their political institutions is one of the pillars of democracy and the decline of trust on institutions is one of the most evident and shared symptoms of the crisis, especially in those contexts where the crisis has been particularly acute. Political trust, or trust in institutions has been measured among European citizens by European Social Survey (ESS) during three time periods, 2008, 2010 and 2012 for 8 countries, namely, Germany, Netherlands, Sweden, Norway, Portugal, Spain, Greece and Ireland. Along this time frame, European societies have witnessed the presence and consequences of the crisis, affecting with different intensity and in different ways their views on the political system and political institutions. We consider trust in institutions as a complex theoretical construct and will be treated as a latent variable using Confirmatory Factor Analysis (CFA). ESS data will be used to test whether the typical indicators for trust in institutions hold across periods and countries, affected differently by the crisis. Our expectation is that the levels of trust, as well as the composition of the latent variable, has experienced clear changes over time, as a consequence of the crisis. In addition to that, we also claim that the correlates impacting trust in institutions are significantly different among countries, depending on the degree in which they have been hit by the crisis. In this paper we will analyze to what extent trust in institutions can be comparable as a construct in different countries with significantly different levels of impact of crisis. In order to study its comparison, tests for measurement invariance of multiple groups (MGCFA) will be used. Latent means will also be studied across time and countries. The main working hypotheses in this paper will be that the levels of political trust are expected to decrease significantly in those countries more affected by the crisis. Moreover, we expect significant differences in the effect of explanatory variables, due to unequal impact of the crisis in countries. We will use Structural Equation Modeling (SEM) methodology in order to analyze and compare the effect of key correlates for trust, such as employment status, satisfaction with economy, satisfaction with life, age, gender, education and observe how the effect of these correlates vary among European societies.

From mixing methods to constructing proxies of indicators: Comparing urban areas in Southeast Asia in qualitative research

Farah Purwaningrum, Institute of Asian Studies Universiti Brunei Darussalam

The paper examines how a mixture of qualitative methods and its analysis can be used as a basis to construct proxies of indicators to enable urban areas to be compared in Southeast Asia. Historically, several urban areas in Southeast Asia were connected due to trading links, such is the case of Malacca Strait connecting ports in Penang, Klang Valley to Batavia/Jakarta. In sociology, urban areas may be referred to as 'meso-sites'. Sociology as a discipline recognises

macro- and micro- levels (see Cicourel 1981, Corcuff 2008). However, ethnographic or organisational context is rarely an explicit source of information for analyses of conversational studies (Cicourel 1981). Quite the contrary, urban ethnographers, or ethnographers in general, see the need to immerse themselves in local setting or community (see Anderson 2002, see also Khosravi 2008). An explicit focus on urban areas may offer alternative ideas of borders as expressed and experienced by local residents as opposed to the kind of borders politically enacted by nation-states (Khosravi 2011). The paper makes such explicit focus on qualitative comparison of urban areas in Southeast Asia. Comparison by qualitative means is a way to grapple with complexity whilst at the same time achieving abstraction by taking into account the context in which a variety of cases is embedded (Palmberger, Gingrich 2014). Building further on the basis of Gingrich's work (see Fox and Gingrich 2002, Gingrich 2002), Koshravi's ethnographic work on 'state of mind' of local residents in light of urban milieu (2008), and Southeast Asian urbanism (Evers and Korff 2000), the paper attempts to explore how three methods namely in-depth semi structured interviews, observation and archival analysis can be utilised as a foundation to construct proxies of indicators for comparison. The author will rely on her empirical data from fieldwork in three urban areas in Southeast Asia from 2010-2015. The areas encompass Jakarta, Indonesia; Bandar Seri Begawan, Brunei Darussalam; and Penang, Malaysia. The proxies of indicators will be constructed on the basis of lived experiences of the residents in the urban areas and archival data. These proxies are akin to variables that can be utilised as points of comparison (see Shtaltovna and Purwaningrum, forthcoming). Limitations and potentials of having such proxies as opposed to other comparative approaches will be discussed. The paper then provides conclusions and recommendations with regard to qualitative comparative approaches.

Cross-national and cross-contextual analysis: Using propensity score matching in the comparison of survey outcomes across countries

Femke De Keulenaer, Ipsos
Ahu Alanya, Ipsos

One of the challenges of survey research in a multi-national, multi-cultural and multi-lingual context is how to select the most optimal question wording among several different possibilities that exist within a language. Different wording alternatives are not always equivalent in terms of comprehensibility - some wordings are more familiar to respondents than others. For instance, more frequent words are recognized and comprehended more quickly than words that have a lower frequency in text corpora (Howes and Solomon 1951). Replacing high-frequency words with low-frequency synonyms reduces response quality as was confirmed by experiments in German (Lenzner et al. 2010), , we conducted an experiment based on a questionnaire that is constructed of a series of PEW internet research questions on the sensitive topic of terrorism. Based on queries on linguistic corpora and lexical databases we developed four versions of the questionnaire where certain wordings were replaced with synonymous terms with different wording frequencies. One version retained the original format from PEW research, another had replaced 12 wordings with more frequent wordings (improved version), while two versions had replaced original wordings with less-frequent wordings: one version had made 15 replacements (worse version) and one 33 replacements (the worst version). The study was carried out on the Survey Monkey Auidence panel where each group had between 720 and 780 respondents. We observed differences for different indicators of response quality: The drop-out rate in the worst version was 20%, while in the other three versions it was about 15%. In addition, there were more DK answers in the worse and the worst version than in the other two. Moreover, fewer respondents enjoyed responding to the questionnaire in the worst version compared to the original PEW questions. There are also some small but not significant differences in response times between the different versions. What we conclude based on our experiment is that low-frequency wordings have a significant effect on response quality only under a certain threshold or when there are several wording changes within a survey question throughout the questionnaire.

Thursday, July 28th, 9:00 a.m. - 10:30 a.m.
Session: Aspects of Data Quality
Chair: Kristen Cibelli-Hibben, University of Michigan
Location: Great Lakes D

Does interview length have an effect on data quality: Evidence from European Social Survey
Laur Lilleoja, Tallinn University
Mare Ainsaar, University of Tartu

The aim of the current study is to evaluate the effect of survey length on the data quality, using European Social Survey (ESS) as an example. The ESS is known for its high standards of data quality (Jowell, et al. 2007), but it has a relatively long questionnaire, which might cause some decline in response quality. ESS questionnaire includes several modules, which some are replaced in every round, but it always includes a Schwartz Portrait Value questionnaire (PVQ-21) in the very end of the questionnaire. Therefore, the PVQ-21 scale provides a suitable personal level indication for response quality in the end of the interview, which can be then related with the overall interview duration. We used three criteria to construct a total survey quality indicator: item non-response, straight lining, and contradictory value assessments to evaluate responses on human value scale. Our analyses indicate to the U-shape relationship between interview length and response quality, as both the fastest and slowest respondents tend to have lower response quality compared with the other respondents. We also found a large cross-cultural variation of proportions of respondents failing in different quality criteria and these results are rather robust across rounds, which refers to the existence of different 'response cultures'. As PVQ-21 is an intensively used value scale, the results hold also an independent methodological importance in the context of value research.

Word frequency effect and response quality: A comparison of four questionnaire versions on a web panel
Ana Slavec, University of Ljubljana
Vasja Vehovar, University of Ljubljana

One of the challenges of survey research in a multi-national, multi-cultural and multi-lingual context is how to select the most optimal question wording among several different possibilities that exist within a language. Different wording alternatives are not always equivalent in terms of comprehensibility - some wordings are more familiar to respondents than others. For instance, more frequent words are recognized and comprehended more quickly than words that have a lower frequency in text corpora (Howes and Solomon 1951). Replacing high-frequency words with low-frequency synonyms reduces response quality as was confirmed by experiments in German (Lenzner et al. 2010), English and Slovenian language (Slavec and Vehovar 2015). However, more research is needed for more cases. In particular, it is not yet known what the thresholds are and how sensitive different response quality indicators are to various degrees of wording changes. To evaluate this, we conducted an experiment based on a questionnaire that is constructed of a series of PEW internet research questions on the sensitive topic of terrorism. Based on queries on linguistic corpora and lexical databases we developed four versions of the questionnaire where certain wordings were replaced with synonymous terms with different wording frequencies. One version retained the original format from PEW research, another had replaced 12 wordings with more frequent wordings (improved version), while two versions had replaced original wordings with less-frequent wordings: one version had made 15 replacements (worse version) and one 33 replacements (the worst version). The study was carried out on the Survey Monkey Auidence panel where each group had between 720 and 780 respondents. We observed differences for different indicators of response quality: The drop-out rate in the worst version was 20%, while in the other three versions it was about 15%. In addition, there were more DK answers in the worse and the worst version than in the other two. Moreover, fewer respondents enjoyed responding to the questionnaire in the worst version compared to the original PEW questions. There are also some small but not significant differences in response times between the different versions. What we conclude based on our experiment is that low-frequency wordings have a significant effect on response quality only

under a certain threshold or when there are several wording changes within a survey question throughout the questionnaire.

When does adding a survey language make sense? Representation bias, response rates, and strategic considerations

Michael Ochsner, FORS, Lausanne and ETH Zurich
Sven E. Hug, ETH Zurich and University of Zurich

With increasing globalization and migration, populations are getting more and more heterogeneous, also regarding languages. Surveying such linguistically and culturally heterogeneous populations is a delicate task needing a thorough translation process and, depending on the mode, multilingual interviewers or a complicated process of assigning interviewers to respondents. Additionally, survey administrators are increasingly under financial pressure, making it difficult to survey a more complex population with less funding. In this presentation, we examine the effects that adding (or removing) languages to a survey has on representation bias or response rates. The first example examines the effect of adding languages. Two surveys were administered in three humanities fields at Swiss universities: English and German literature studies, and art history: The first survey was administered in English and German, thus covering the language of the first two fields, including the language of the majority in Switzerland. The second was administered adding a third and fourth language, namely French and Italian, two Swiss national languages that are at the same time very important scholarly languages in the third subject field. We examine the representation regarding language region (Swiss and French part of Switzerland) and subject field and examine the mother tongue as well as the language chosen to fill in the questionnaire. The second example examines reducing languages using a general population survey of a Swiss city that was to date administered as a telephone survey in multiple languages. Due to budget constraints and especially severe drops in the response rates over time due to decreasing phone coverage, a single-language web/paper mixed-mode experiment was conducted. We find that using a more inclusive mixed-mode design can compensate for some representation bias when reducing languages. The two examples shed light on the advantages and disadvantages of using multiple survey languages. It also reveals some practical implications for deciding how many and which languages to choose when administering a survey: strategic considerations have to be made since ethical and political issues (inclusion of minorities), methodological effects (change of mode to compensate), representation bias (are minorities large enough to make a difference), language competence in the surveyed population, and financial constraints have to be balanced out.

Thursday, July 28th, 9:00 a.m. - 10:30 a.m.
Session: Survey Mode

Chair: Julie de Jong, University of Michigan
Location: Michigan Ballroom I

Assessing the feasibility of using SMS Surveys in Eastern Europe to conduct a cross-country comparison of attitudes with regard to Europe's migration crisis.

Sara Gjoshevska, Sample Solutions
Carsten Broich, Sample Solutions

Cross country data collection is a complex process however technology has enabled worldwide research to embrace various interviewing channels that can facilitate this process. According to the FocusVision 2014 Annual MR Technology Report, trends in research technology show a high usage of more high-tech modes but also lower volume/minor modes such as SMS surveys are now on the rise. SMS is a minor mode of data collection administered through a self-completion mode and it gives respondents time to reflect and they may give fewer rounded answers and more disclosure of socially undesirable behaviors and views differing from the general public opinion, making it suitable for gathering data on sensitive issues such as migration. Immigration is a highly researched topic through opinion polls and surveys, the latest Standard Eurobarometer polling (Autumn 2015) by the European Commission

showed that Immigration ranks among the most important issues facing the EU currently. This research will attempt to show if SMS Surveys are a practical method for collection of cross country information on public attitudes aimed at immigration, and the various dimensions of economic, public and private life that individuals feel are affected by immigration. The design and implementation of the cross-country survey methods for this research face multiple factors such as ensuring the sampling procedures are consistent across nations with complete surveys N=500 per country and all distributed concurrently. Furthermore, cross-national research design requires specific survey design such as keeping the survey questionnaires as consistent as possible across all countries by overseeing the survey translation process. To maintain the quality of the survey implementation three stages are implemented: a pilot study, the main data collection stage and a follow-up study with the respondents. For comparability reasons, the objective of the study was to use three countries, specifically Poland and Romania as part of the European Union (EU) and The Republic of Macedonia as a non-EU country, hereby highlighting the various benefits and constraints of using SMS surveys and the differences between the three countries thus comparing demographic data in order to examine the differences in attitudes and the comparability across the three countries pertaining to the current migration crisis in Europe.

The use of SMS pre-notices for CATI interviews and their impact on response rates, and non-respondents by age and gender

Sofia Kluch, Gallup
Ken Kluch, Gallup

Using Gallup World Poll, this study examined the use of SMS pre-notices for a nationwide CATI study in Sweden. Based on sample data, both response and non-response of potential participants was examined from a randomly selected general population mobile and landline sample in Sweden. Drawing on public databases, the Swedish general population sample provides detailed demographic information. Different response and refusals patterns were explored by demographics to better understand non-response and how the use of a text message pre-notice impacted both the likelihood of cooperation and the form of non-response. The sample data was compared to final study completes and Sweden census data to evaluate the representativeness of the sample, completed interviews, and outcomes of each call attempt on the total sample. Earlier World Poll research in Norway indicated that some segments of the population non-respond actively, that is they choose to not participate, typically via a direct refusal. In contrast, other segments of the population non-respond passively , that is they never explicitly stated they would not participate, but their non-respond either through sending the call to an answering machine, or simply not answering the call. The current study seeks to determine if the use of text messaged pre-notices reduces the occurrence of passive non-response by increasing awareness of the purpose for the call and providing the number and caller ID details for the interviewer. The use of SMS pre-notices did yield an increased response rate and cooperation rate and reduction in the refusal rate. The sample is currently under examination for demographic differences. Among the population not receiving the SMS pre-notice, we do not expect demographic differences in non-response. Among the respondents receiving the SMS pre-notice message, we expect a significant difference in age, with younger potential respondents being less likely to passively non-respond to the survey.

Quality of leave behind paper questionnaire: Comparison of three national representative surveys to approach errors in sampling and attitudinal variance

Alexandre Pollien, FORS
Dominique Joye, University of Lausane
Marlene Sapin, FORS

Quality assurance and control is a crucial component of national and cross-cultural surveys. For many researchers in the field of survey methodology the quality of a survey is approached in the dominant total survey error paradigm. While the sampling variance is assessable in most probability sample survey, other error components are much more difficult to gauge without alteration of the whole survey design, notably the variance in the attitudinal structure. In

this paper, we compare three Swiss representative face-to-face surveys of the MOSAiCH (Measurement and Observation of Social Attitudes in Switzerland) programme, including two ISSP (International Social Survey Programme) modules at each edition), in which a paper questionnaire related to the main topics of the face-to-face survey was left behind. The survey design was alike in these three editions, but the topic of the survey and of the supplemental drop-off questionnaire was different: the main topic of MOSAiCH 2009 was around religion, the one of 2013 on family, while the subject of MOSAiCH 2015 focused on work issues. Such comparison permits to identify error in representativeness in the leave-behind questionnaire due to the attrition of unavailable or jaded respondents. However, the attrition can also be related to attitudes towards the topics, implying discrepancies in measurement depending of the questionnaire considered (main or leave-behind). This paper intends to show that somewhat close surveys designs give rise to different quality of data, when considering the sampling variance and attitudinal propensity to answer. Looking at three surveys contexts as whole entities leads to questioning about comparative issues. Sampling variance and attitudinal discrepancy related to the survey topic, as components of the total survey error, interrelate differently in each survey context. This paper accounting for surveys as contextual entities, stresses the interrelation of the survey topics with survey design.

Finding the Best Fit: An Comparison of CAWI testing to CATI-trended data for general population research in Germany, Italy, Denmark, Sweden, and the United States

Sofia Kluch, Gallup
Ken Kluch, Gallup

With declining response rates, researchers are exploring a variety of multimode methods of collecting data. This challenge is compounded in cases where researchers are conducting nationally representative studies. The Gallup World Poll has been fielded via phone annually in Denmark, Italy, Germany, Sweden, and the US since 2006. In 2015, Gallup completed three CAWI tests to determine which, if any, survey items are sensitive to mode differences. Three forms of the original CATI instrument were designed and fielded to two probability based panel samples. One sample was provided by the in-country vendor and the second sample included respondents who had previously completed a Gallup World Poll survey in 2014 or 2015 and that agreed to future recontact. The three CATI forms differed in response options and instructions. Form 1 provided a don't know and refused option for every question and required answers for each question on the instrument. Form 2 removed the don't know and refused options and included explicit instructions at the beginning of the survey that items could be skipped if necessary, but all responses were requested. Form 3 followed the same format as form 2, but also included additional instructions and clarifications from the CATI version, such as interviewer instructions, prompts, probes, or definitions for potentially problematic items. The additional instructions were used in form 3 to determine if additional content would impact both the type of response and any potential non-response in cases where the respondent wanted more information. This paper presents data collected in all four countries, including the items that are most sensitive to mode effect, the format that best controls for mode effects and differences in the type of respondents in each of these scenarios.

Thursday, July 28th, 9:00 a.m. - 10:30 a.m.

Session: Open Panel Alliance

Chair: Michael Bosnjak, Marcel Das, Arie Kapteyn, GESIS - Leibniz Institute for the Social Sciences, CentERdata / Tilburg University, Center for Economic and Social Research, USC
Location: Great Lakes E

Understanding America Study

Arie Kapteyn, Center for Economic and Social Research, USC

The Understanding America Study (UAS) is a new household panel maintained by the Center for Economic and Social Research (CESR) at USC, comprising approximately 2,500 households representing the entire U.S. (it is planned to grow to 6,000 by the end of Summer 2016). The study is an 'Internet Panel,' which means that respondents answer

our surveys on a computer, tablet, or smartphone, wherever they are and whenever it is convenient for them to participate. From the viewpoint of representativeness, it is important to note that the panel is recruited through address-based sampling (ABS). Anyone willing to participate who does not have a computer or Internet access is provided with a tablet and broadband Internet. Increasingly, address based sampling is becoming the method of choice for building population representative panels, since other methods such as Random Digit Dialing (RDD) are increasingly facing challenges as a result of the change in telephone use in the population. Panel members answer questions in waves about once or twice a month. Surveys are restricted to about 30 minutes per interview, but since all data can be linked across waves, a large amount of information is available about panel members, including demographics, financial behavior and financial literacy, cognitive capability, and personality. Of specific interest is the fact that all panel members answer the full survey instrument of the HRS, which contains some 90 minutes worth of interview time (administered over a number of separate sessions to stay within the 30 minute limit per survey) with information on health, income, assets, labor market position, and expectations. Respondents receive compensation for their time spent answering questions at a rate of $20 per 30 minutes of interview time. Annual attrition rates are modest (on the order of 6-7% per year). Data are made available to the research community free of charge and anyone wanting to collect new data in the UAS can combine the newly collected data with the information already available. The recruiting response rate of the UAS is approximately 20%, higher than any other probability Internet panel in the U.S. The core of the UAS team joined CESR from RAND where the team built and managed the RAND American Life Panel and developed the survey software package MMIC.

LISS Panel: Collecting representative data for the Netherlands

Marcel Das, CentERdata / Tilburg University

In this presentation, the LISS Panel, Dutch member of the Open Panel Alliance, will be presented. The LISS panel consists of about 7,000 individuals and is representative of the Dutch-speaking population in the Netherlands. The panel is based on a probability sample drawn by Statistics Netherlands from population registers. Respondents answer interviews over the Internet monthly. Households that could not otherwise participate are given a computer and broadband Internet access. The LISS panel is the central source of the MESS facility. In 2006, the Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences (MESS) was started as one of five large research infrastructures funded by the Dutch government. MESS is an innovative data collection facility intended to boost and integrate research in various disciplines, such as social sciences, life sciences, and behavioral sciences. Other key elements of the MESS project are: (1) a longitudinal core questionnaire and experimental modules proposed by researchers from all over the world. The core questionnaire, designed with assistance from international experts in the relevant fields, contains questions on topics such as health, work, income, education, ethnicity, political opinion, values, norms, and personality. Designed to follow changes over the life course of individuals and households, it is repeated annually. (2) innovative forms of data collection. Besides traditional questionnaires, the facility accommodates innovative ways of asking survey questions, e.g., exploiting visual tools on the screen or collecting data in other ways than through survey questions. This includes various new communication and measurement devices like smartphones with GPS, as well as devices to measure biomarkers such as weight, bioelectrical impedance, and physical activity levels. These tools allow for much more accurate and cost-effective measurement and experimentation in large representative samples than was possible in the past, leading to richer and better data on many domains of people's lives. (3) linking with administrative data. Administrative data on, for example, income, assets, and pensions archived at Statistics Netherlands can be linked with LISS panel data. Statistics Netherlands provides a remote access facility through which the linked data can be used. Powerful elements of MESS are its open access and its population representativeness, providing an environment for cross-disciplinary studies and experiments on a wide array of topics and using advanced measurement devices. Rich background information on many aspects of the respondents' lives is collected or updated each year and made available free of charge.

GESIS Panel: Collecting representative data for Germany

Michael Bosnjak, GESIS - Leibniz Institute for the Social Sciences

In this presentation, the GESIS Panel, German member of the Open Panel Alliance, will be presented. The GESIS Panel is a probability-based mixed-mode access panel infrastructure operated by GESIS - Leibniz-Institute for the Social Sciences in Mannheim, offering the social science community a unique opportunity to collect survey data from a representative sample of the German population free of charge. Data collected within the GESIS Panel can be used by academic researchers free of charge, e.g. to conduct secondary research. By the end of the recruiting phase in February 2014, the GESIS Panel encompassed almost 5000 panelists. The omnibus survey waves take place on a bi-monthly basis, each encompassing about 20 minutes and are split up into two self-administered survey modes (online, offline). About 65% of the panelists participate online (web-based surveys) and about 35% of the panelists participate offline (by mail). Each survey wave consists of two major parts: About fifteen minutes of survey time are reserved for submitted studies. Fielded panel studies from external researchers will have undergone a peer-review process. The second part of each survey wave (about five minutes of interviewing time) is reserved for longitudinal core study topics. One aim of the GESIS Panel Longitudinal Core Study is to measure frequently demanded characteristics beyond demographics, such as, for instance, personality and human values, political behavior and orientations, well-being and quality-of-life, environmental attitudes and behavior, and information/communication technology usage. Moreover, a second aim of the GESIS Panel Longitudinal Core Study is to assess and to control for data quality by measuring concepts such as survey participation evaluations, survey mode habits and preferences, and by including selected items from other benchmark surveys (e.g., German micro-census, ESS, ALLBUS, ISSP).

Thursday, July 28th, 9:00 a.m. - 10:30 a.m.
Session: Questionnaire Design

Chair: Alisú Schoua-Glusberg, Research Support Services
Location: Great Lakes A/B

Design and cultural adaptation of a 22-country tool for collecting comparative data on poverty level

Alisú Schoua-Glusberg, Research Support Services
Katherine Kenward, Research Support Services
Anahit Tevosyan, FINCA

FINCA is a microfinance institution with the mission to alleviate poverty through lasting solutions that help people build assets, create jobs and raise their standard of living. Between 2002 and 2012 FINCA utilized several versions of a Client Assessment Tool (FCAT) to measure and report on client living standards and poverty levels. Since its introduction, questions were added to the tool until it became too long to be practically implemented on an ongoing basis. In 2008 a new tool (the Standard of Living Tool or SLT) was created including 16 of the most important questions from the longer instrument. The tool was further refined in 2011, however feedback from field survey managers showed a need for further refinements. Statistical analysis of the data also revealed some questions were not leading to reliable results. Finally, a literature review on household poverty surveys revealed several weaknesses in the expenditure measurement methodology that required addressing. The most appropriate and unbiased method to measure consumption poverty is via Living Standard Measurement Study (LSMS) surveys, but these were too costly for FINCA's small research unit. FINCA then set to build an abbreviated consumption survey, followed the path proposed by Lanjouw and Lanjouw (1996 and 2001) by analyzing recently implemented full-scale household expenditure/budget surveys in every FINCA country and selecting the consumption items that had strong correlation with poverty. Research Support Services worked with FINCA on instrument development. The work included improving the survey design, construction and question sequencing and phrasing, country specificity issues, reviewing translations, and developing a surveyor's manual. The mandate was to build a new survey instrument that would contain the optimum type and number of questions to accurately measure client living standards while taking into account the operating environment in which the microlender operates, the need to measure poverty outreach on an annual basis, and resource constraints. First, we designed a modular instrument of a generic nature that could be adapted to each country while preserving comparability. Then, with data from prior years about relevant local goods and assets in wealth and consumption measurement, and consulting with local field experts, code lists were

built for each country, maintaining consistent codes for each item across countries. This presentation will highlight the steps followed, challenges faced, and will discuss issues found during data collection and analysis that feed back into instrument refinements. Translation and other language issues will be part of the discussion.

Do response effects generalize across countries?

Henning Silber, GESIS
Annelies Blom, University of Mannheim
Tobias Stark, Utrecht University
Jon A. Krosnick, Stanford University

A great deal of literature demonstrates that, in surveys in the U.S., changes in the design and wording of survey questions produces predictable effects on the distributions of answers. Such effects are called response order effects, question order effects, no-opinion filter effects, acquiescence response bias effects, and others. In this project, we implemented the same experiments in probability samples surveys in 13 countries (2 each in the U.S. and Germany). This paper will present the results of a set of these experiments and compare them across countries. In some instances, the same response effect appears across countries. In other instances, effects vary considerably across countries, but in ways that make sense when taking into account variation in the conditions necessary to observe an effect. These findings have important implications regarding the development and application of optimal principles of questionnaire design across countries and languages.

Sequence matters: The impact of the order of probes on response quality, motivation of respondents, and answer content

Katharina Meitinger, GESIS - Leibniz Institute for the Social Sciences
Michael Braun, GESIS - Leibniz Institute for the Social Sciences
Dorothee Behr, GESIS - Leibniz Institute for the Social Sciences

Due to the growing significance of international studies, the need for tools to assess the equivalence of items in international surveys is pressing. Online probing is a powerful tool to identify the causes of non-equivalence by incorporating probing techniques from cognitive interviewing in cross-national Web surveys. So far, online probing applies three different probe types - category selection probes, specific probes, and comprehension probes - to inquire about different aspects of an item. Previous research mostly asked one probe type per item but in some situations it might be preferable to test potentially troublesome items with multiple probe types. However, we miss empirical evidence on whether the order of probe types has an impact on the response quality and the answer content. This presentation will report the results of a Web experiment conducted with 1,354 respondents from Germany, USA, Mexico, Spain, and Great Britain in June 2014. The participants were selected from a non-probability online access panel using quota for age, gender, and education. Each of the experimental groups received three probes (category selection, specific, and comprehension probe) on separate screens for one closed item (How important is it that people convicted of serious crimes lose their citizen rights?) from the ISSP item battery on people's rights in a democracy. The order of probes was changed in each condition. This presentation will address four research questions: First, is the quality of probe responses affected by the sequence in which the different probe types are being asked? We gauged response quality by the length of probe responses, incidences of probe non-response, and incidences of probe mismatches (e.g., a respondent answering a specific probe with a response to a category selection probe). Second, does the motivation of respondents differ across split conditions? In particular, do respondents show signs of a reduced motivation and do they satisfice? Third, does the order of probes have an impact on the answer content? For example, a previous probe answer could frame the response to the following probe. Finally, does the probe order influence the respondents from the five countries differently?

Thursday, July 28th, 9:00 a.m. - 10:30 a.m.
Session: Questionnaire Design and Testing 1

Chair: Ana Villar, City University London and Sarah Butt, City University London
Location: Huron

*** Overview of questionnaire design and testing**

Ana Villar, City University London
Brita Dorer, GESIS-Leibninz-Institute for the Social Sciences

*** Setting up the cognitive interview task for Spanish, Chinese and Korean-speaking participants: How is the introduction best tailored to different groups?**

Hyunjoo Park, RTI International
Patricia Goerman, U.S. Census Bureau

Cognitive interviewing is a method used to identify problematic survey questions by asking research participants to report what they are thinking, either while answering survey questions or retrospectively. It is a popular method to pretest the quality of questionnaires as it allows in-depth evaluation of how participants mentally process survey questions. In recent years, the body of empirical research on cognitive methods has increased. Some studies have discussed difficulties in applying standard cognitive interview methods to different populations, such as participants with low educational attainment or income levels, regional differences and linguistic differences, even among English speakers within the U.S. Due to rapidly growing non-English speaking populations in the U.S., cognitive interviewing of survey translations has become a standard practice. However, the aforementioned difficulties in applying standard cognitive interview procedures become more complex when using this method to test survey translations. Previous research has reported that participants who speak little or no English often exhibit difficulties providing adequate answers in cognitive interviews because they are unfamiliar with surveys in general or have limited exposure to the English language and mainstream American culture (Coronado & Earle, 2002; Kissam, Herrera, & Makamoto, 1993; Pan, 2004; Pasick, Stewart, Bird, & Donofrio, 2001). Due to the differences in communication styles across languages and cultures, these difficulties are somewhat expected. For example, communication in English relies more on the facts of the message than on context and background information, and is characterized as low context. The actual content of the message is more important than when, how, and by whom it is expressed. In contrast, speakers of high-context languages, such as Chinese and Korean, rely heavily on context and interpersonal cues (Hall & Hall, 1987). Politeness is highly valued in Asian and Hispanic cultures, and silence and acquiescence are often interpreted as polite behavior (Javeline, 1999). Since cognitive interviewing has historically been implemented in the communicative norms of English and Western cultures, where directness and openness is a preferred communication style (Pan et al., 2010), these cross-cultural differences in communication style add another layer of complexity to implementing standard cognitive interviewing methodology in non-English interviews. However, very little research has been conducted on this issue. One question in particular is how best to explain the cognitive interview process and purpose to participants prior to conducting an interview and whether this should be done differently across language and cultural groups. If respondents understand the purpose of the cognitive interview and the probing questions, they may be more able and willing to provide responses that will assist researchers in evaluating the survey questions. This paper aims to fill a gap in the literature through development and evaluation of introductions that explain the cognitive interview task to non-English speaking participants. We present preliminary evidence from two studies with Chinese, Korean, and Spanish-speakers, where traditional and enhanced introductory scripts were used to introduce the cognitive interview to participants. In a project focused on Chinese and Korean speakers, we used three different sets of scripts in practice sessions at the beginning of the interviews. First, we adapted a practice question described in Willis (2005) and Goerman (2006) and had participants answer How many windows are there in your home? However, we found that this was not an ideal practice question for our study since it was originally designed to induce the respondent to think aloud, while most of the cognitive tasks in our interview involved meaning-oriented probes (i.e. What do you think they mean by term xxx?) designed to evaluate whether participants

comprehended the question wording. Secondly, we asked the participants to answer an open-ended question what is your favorite season? followed by three probes: a procedural probe; an interpretive probe and a paraphrasing probe. Finally, we asked the participants to fill out an intentionally damaged survey question with response options on paper to simulate the real interviewing task using the same survey administration mode with two types (interpretive and evaluative) of probes. Then, we explained the reasons for such practices. In a Spanish-language cognitive interview project, we examined the effects of two interviewing techniques: 1) standardized interviews following the same procedures as typical U.S. English-language cognitive interviews, with direct translation of an English interview protocol; and 2) experimental interviews allowing for variation of introductory statements and conversations. In this study, the interviewers completed the "How many windows" question aloud themselves rather than asking participants to do so, demonstrating both how the question would be asked and how the think aloud exercise would work in practice. Interviewers then presented participants with a sample meaning oriented probe saying I might then ask you a question that sounds a bit strange, like what does the word window mean to you? Interviewers then answered this question themselves and gave respondents an example to illustrate that different people might respond differently to the question, which would help us to evaluate the effectiveness of the survey question. In this paper, we will discuss the qualitative findings from these two projects in terms of perceived success in tailoring the introductory statements to the different language groups. In the Spanish project we coded interactions to identify discomfort and whether or not the respondent s answer was useful in evaluating the survey question. In the Chinese and Korean project, we modified introductory scripts in an iterative process over different rounds of testing based on the interviewers perception of what worked best. Based on evidence from our studies, we will propose baseline introductory procedures and scripts for use with Chinese, Korean and Spanish speaking respondents. We will also describe some of the major research questions and possible designs for future cognitive testing projects to expand this line of research.

Measuring alcohol consumption in the ESS: Coordinating question adaptation across 23 countries

Ana Villar, City University London
Lizzy Winstone, City University London

Virtually all cross-national surveys face the task of question adaptation. Adapting survey questions for different subpopulations is sometimes a necessary but often complicated process, and researchers lack both guidelines on how to carry out adaptation and actual examples of the outcome of these processes. This paper intends to reduce this gap by describing and evaluating the adaptation of questions on alcohol consumption designed for Round 7 of the European Social Survey (ESS, 2012). When it comes to measuring alcohol consumption, differences on types of alcoholic drinks, on the containers in which they are served, and on the average alcohol volume in drinks that have similar names can lead to measurement differences that render comparisons across countries impossible. For this study, we considered adopting approaches to measuring alcohol consumption in existing cross-national surveys. However, none of the approaches we found proved suitable: some involved a number of questions that surpassed the maximum available space in the questionnaire; others were found to be too difficult by respondents, who needed help from the interviewer in our pilot study; others did not seem to take into account differences in typical container size and therefore resulted in measures that were deemed not comparable. More importantly, for most of the surveys, only the prototype version of the showcards was available from documentation, which made the adaptation of the showcards still necessary for implementation across the 23 countries that participated in Round of the ESS. Therefore, a new set of items was developed to measure alcohol consumption using four items, trying to simplify the respondent's task as much as possible. These items tried to match the way respondents think about alcohol consumption. The approach called for country-specific categories to define different quantities of alcohol and for images that illustrated specific alcoholic drinks as well as binge drinking amounts, with planned harmonisation of responses into total grams of alcohol at the data processing stage. The aim of this approach was to obtain measures that were as comparable across countries as possible and that were easy for respondents to answer. This paper attempts to contribute to improving understanding of challenges behind adaptation efforts. To do this, we will describe the coordination of this adaptation procedure, show examples of the actual questions and showcards fielded in Round 7 of the ESS, and evaluate the approach.

Developing a cross-national measure of ancestry for use on the European Social Survey

Sarah Butt, City University London
Anthony Heath, Nuffield College, Oxford
Silke Schneider, GESIS

A person's ethnic or socio-cultural background has been shown to be an important predictor of a range of social attitudes and behaviours. Ideally, therefore, we want to capture such information alongside other demographic variables in social surveys. However, gathering information about people's socio-cultural origins as part of a cross-national survey is complicated, not least because of the need to capture complex variation in national, ethnic and other cultural groupings prevalent across countries. The European Social Survey (ESS) recently trialled an approach to collecting data on socio-cultural origins based on a measure of respondents' self-reported ancestry developed by the Australian Bureau of Statistics and fielded as part of the Australian Census. Whilst the Australian measure provided a good starting point, two types of adaptation to improve conceptual coverage were considered necessary before the item could be fielded as part of a pan-European survey. First, the Australian-focused codeframe, the Australian Standard Classification of Cultural and Ethnic Groups (ASCCEG) needed to be adapted for a European context e.g. by allowing for greater differentiation between different types of European ancestry. Second, country-specific showcards were developed to provide respondents with comparably salient stimuli for generating relevant responses. Following the first round of data collection for the ancestry item in ESS Round 7 (2014/15) a thorough evaluation of both the instrument development process and the survey data collected has been carried out. Preliminary findings suggest that the item worked well across ESS countries and generated meaningful data on respondents' socio-cultural origins. The harmonised code-frame and country-specific showcards developed during the adaptation process are judged largely fit for purpose. However, the evaluation also highlights a number of ways in which they could be improved for future rounds, especially in relation to capturing sub-national groups. In this paper we reflect on the adaptation process followed when developing an ancestry item for the ESS. We first describe the approach taken to developing a new European Standard Classification of Cultural and Ethnic Groups and the process of consulting on and documenting country-specific adaptations. Drawing on the evaluation findings we then highlight aspects of the adaptation process that appeared to work particularly well and those that worked less well before going on to suggest some possible lessons for future adaptation exercises on the ESS and similar cross-national surveys.

Thursday, July 28th, 11:00 a.m. - 12:30 p.m.
Session: Analysis Methods and Tools 2

Chair: Timothy P. Johnson, University of Illinois at Chicago
Location: Michigan Ballroom II

A multilevel approach to understanding contextual effects on youth's literacy and learning activities: Two examples from PISA 2009 and PIAAC 2013

Suehye Kim, UNESCO Institute for Lifelong Learning

This presentation aims to investigate contextual effects on outcomes in youth's literacy and learning activities using two international-large scale assessment studies, the Programme for International Student Assessment (PISA) 2009 study and the Programme Internatioanl Assessment for Adult Competencies (PIAAC) 2013 study conducted by Organisation for Economic Co-operation and Development (OECD). PISA 2009 study places a particular interest on 15-year-old youth's reading competence, by assessing not only reading knowledge and skills, but also their attitudes and tastes in reading. Focusing on reading, it provides various indicators of youth's literacy and reading activities. As a life-span approach, the PIAAC also measures young adults (25 and 34 years of age)' literacy skills defined as abilities to understand, to evaluate, to use and to engage with written texts in order to participate in society. Given that further learning is essential for young adults to live well in the age of information and communication technology, non-

formal adult learning activities are investigated in the PIAAC. A larger socio-economic and cultural context shapes literacy and learning activities. The importance of culturally instilled beliefs differs by cultures, so these non-educational effects need to be considered in cross-national comparative studies (Meyer & Schiller, 2013). Identifying national variations may have an advantage over research restricted to individual countries. For this, I will adopt the Hofstede cultural dimensions of national culture with two measurements as follows. Conceptually, Power Distance (PDI) indicates the less powerful members of society accept that power is distributed unequally, while Individualism versus Collectivism (IDV) describes the degree to which a culture relies on and has allegiance to the self or the group. Since these two dimensions are highly and significantly correlated, I will suggest using a new indicator which indicates the relative degree of Individualism over Power Distance. Also, I will gauge to what extents each country emphasizes achievement and ambition by Masculinity versus Femininity (MAS) index, which indicates the degree to which a culture values such more masculine behaviors as assertiveness, achievement, and acquisition of wealth versus caring for others, social supports and quality of life. Considering the observed national variations attribute to different sociocultural environments, this contribution will demonstrate an analytical technique that captures the effects of cultural characteristics on the literacy achievement among OECD countries.

How to record events' data on country level for cross-national comparisons? Example of the ESS Round 6 and 7 in Poland

Teresa Zmijewska-Jedrzejczyk, University of Warsaw / Polish Academy of Sciences
Danuta Przepiórkowska, University of Warsaw / GESIS

When analyzing data from longitudinal, cross-national surveys one must be aware of country-level differences. A particular part of context data comprises events such as terrorist attacks, financial crises, public affairs or political scandals. Either of them might influence individual opinions in different ways across countries. The necessity of measuring such contextual effects of events had been pointed out in 2015 at the comment to the document "The implementation of the ten United Nations Fundamental Principles of Official Statistics" published Statistical Journal of the IA (Ref.). Author of that comment also mentioned that in the European Social Survey (ESS) problem had been addressed and innovative approaches to collect media claims were implemented in 2012. Data from ESS Round 6. from 2012 and 7. from 2015 in Poland provide an evidence that current approach has a specific, conceptual limitations: (1) events present in the public sphere but not reported in quality broadsheets (e.g. present only in tabloids) are omitted; (2) only claims literally connected to the topics of the ESS questionnaire are included; (3) events not occurring during the fieldwork period are not included. Then, using the adapted model of Coleman's scheme (Raub, Buskens, Van Assen 2011) a new approach to measure the contextual effects will be presented.

* Addressing equivalence and bias in cross-cultural survey research within a mixed methods framework

Jose-Luis Padilla, University of Granada, Spain
Isabel Benitez, University of Granada, Spain
Fons van Vijver, Tilburg University

Detection of potential sources of bias in cross-cultural research is cursory in the best beyond the growing attention paid to the translation and pretesting phases. The lack of a comprehensive and systematic approach is particularly serious when different linguistic versions of psychological or health scales are included in survey questionnaires. Frequently, survey researchers resort to previous validation studies of scales taking for granted their psychometric properties or considered them as "permanent". The objective of the paper is twofold: a) to present shortly an integrated approach to bias in cross-cultural survey research; and b) to illustrate how such approach can be applied within a mixed research framework combining quantitative and qualitative methods. We review current consensus about how to reach integration through the three main integration levels: design, method, and interpretation and reporting. Then, we identify the approaches in each level that can be more useful for studies aimed to detecting bias in cross-cultural research. To illustrate the integrated approaches to bias detection, we present several real cases of international survey research projects in the health, quality-of-life and educational fields.

Thursday, July 28th, 11:00 a.m. - 12:30 p.m.
Session: Data Collection Challenges and Approaches

Chair: Christof Wolf, GESIS
Location: Great Lakes A/B

* Linking auxiliary data to survey data, ethical and legal challenges in Europe and the US

Kirstine Kolsrud, Norwegian Social Science Data Services (NSD)
Katrine Utaaker Segadal, Norwegian Social Science Data Services (NSD)

Auxiliary data in a survey context is in a broad sense all forms of data that are not collected directly from the respondents, and can include a vast variety of information. Typical sources of auxiliary data are information from the sampling frame, census data, administrative data or data from private data collectors (Kreuter 2013). Such additional information could for instance consist of neighbourhood contextual information such as population density, crime rates and poverty, government records on tax collection, pension or welfare benefit computations or public available material from social media.

The interest in the use of auxiliary data in survey research has notably increased over the last years. Judging from the literature this seems to a large extent to be driven by the theoretical promise of auxiliary data to overcome the challenge of nonresponse (Krueger and West 2014). However, the increased demand for auxiliary data to be applied in survey research is naturally also ascribed to new and improved opportunities to access and utilize these kinds of data in combination with survey data. It is further argued that such data can enhance research quality by extending the possibilities of analysis and also by obtaining more reliable survey data, whereas high reliability is one of the main principles of integrity in science. Auxiliary data is thus highly relevant both for methodological and substantive survey research.

While confidentiality including data security is one of the cornerstones of data protection, informed consent stands out as another key principle in research ethics to protect the interest of research participants. However, both confidentiality and informed consent in relation to complex and increasing use of data linkage might bring about several issues.

Legislation and practice in the data protection field in Europe has until now been highly fragmented, particularly in a cross country setting. However, a new General Data Protection Regulation is scheduled for formally adoption by the EU Parliament and the Council by March this year and within two years from that date, the Regulation should be implemented in all EU and EEA countries. The promise of the Regulation is to ensure a consistent and high level of personal data protection to provide legal certainty and trust, which should be equivalent in all Member States.

In the US, the government is currently in the process of reconsidering the Common Rule. Major changes are related to tighten the rules of consent, make the review system more effective and establish mandatory data security standards.

The presentation will focus on ethical dilemmas and legal regulations, especially with regards to confidentiality and (exemption from) consent, in light of both current practice and new framework conditions for research within Europe and the US.

* Multinational event history calendar interviewing

Yfke Ongena, University of Groningen
Marieke Haan, University of Groningen
Wil Dijkstra, University of Groningen

In our paper we discuss how we developed a training and monitoring scheme adapted to Event History Calendar (EHC) interviews in a large-scale survey on smoking history. The study was conducted in five different European countries (France, Germany, Greece, Italy and Slovenia). We also report on differences in interviewer behavior, based on behavior coding data of 914 interviews, administered by 116 different interviewers from these five countries. EHC interviews have been developed to improve respondent's recall of retrospective information, and are becoming a more and more popular replacement of conventional (standardized) interviews. The EHC is usually a matrix with column headings indicating calendar years and/or months, and rows representing different life domains (i.e., residence, marriage, employment etc.). Life events entered in the calendar serve as cues, enhancing the retrieval of other events. By inspecting well remembered events from previous domains (e.g. moving houses or changing jobs), the interviewer can help the respondent to retrieve other events, with cross-referencing probes like Did you quit smoking before or after you moved to Amsterdam? The effects of EHC methods on data quality have been examined already in several studies (for an overview, see Glasner & Van der Vaart 2007), but not much research has been done with regard to interviewer effects and effective interviewer training in administering EHCs. In addition, EHC studies are usually conducted within one country. No studies have examined the feasibility of EHC interviews in multiple countries, in multiple languages. In our study, an electronic EHC was developed and two laptops were used to administer the interview. On the first laptop, the interviewer could fill out the EHC and switch to screens with question wordings for each domain. The second laptop (which was connected to the first) was used to show the respondents the EHC being filled out with their life events. It is assumed that this helps them in remembering dates and events. In addition, the interviewer could show the response alternatives for closed-ended questions on the respondent's laptop. The interviews were audio-recorded, using the interviewer laptop as a recording device, and a semi-automatic behavior coding procedure was implemented. The so-called log file keeps track of all mouse clicks and key presses by the interviewer, and thus provides additional information about the interviewer performance. For example, in case a question is accompanied by response alternatives, the log file tells us whether or not the interviewer showed these alternatives on the respondent's laptop. The audio-files and log files were sent to coders in the different countries, who behavior coded parts of the interviews. Based on information from the log file, the behavior coding software made a selection of those parts of the EHC that were subjected to behavioral coding. For example, if it appears from the log file that a particular question-text window was not opened by the interviewer, the program decides to check question reading for that question. Question texts were available in different languages in the program, to allow the coder to compare exactly the scripted texts of questions and response alternatives with the texts the interviewer read to the respondent. The behavioral coding data concerns both standardized behavior like question wording, and behaviors that are EHC specific, such as cueing and cross-checking, or probing for changes. Also data entry errors are checked, by comparing what the respondents report according to the audio-recorded interview, with what is entered by the interviewer. The first, second, fifth, and tenth interview of each interviewer, and thereafter ten percent randomly selected interviews of each interviewer, were subjected to behavioral coding. Research has shown that in general interviewer performance tends to decrease over time, but is improved again after interviewers obtain feedback about their performance (Van der Zouwen and Dijkstra, 1995). Therefore, interviewers received feedback about their performance, and recommendations about how to improve their behavior. This feedback was based on the behavior coding data of at least two interviews that the interviewer had conducted between two rounds of feedback. The results showed that it is important to monitor interviewers throughout the whole fieldwork period. We found that the interviewer behavior immediately after training was better than when interviewers had administered several interviews without having obtained feedback about their performance. With each round of feedback, interviewer behavior significantly improved. A multilevel analysis, in which interviewer variance was taken into account, showed no significant impact of countries for standardized behaviors like question reading and probing for alternatives. This demonstrates our training program was successful in equalizing standardized behavior across the five countries. However, countries did differ significantly with respect to the extent to which interviewers applied the Event History Calendar principles. In the paper we will fully describe the results of multivariate analyses, comparing interviewer effects across the countries, also taking into account interviewers and respondents demographic characteristics such as age and gender. Since the study was designed as a case-control study (on smoking history), we will also compare cases and controls. Cases were lung-cancer patients recruited from a hospital, controls were patients diagnosed with another disease, recruited from the same hospital. For

interviewers, additional information was available on their certification history (i.e., were they certified as interviewer immediately after their first successful interview, or did it take several rounds of feedback before they were certified as interviewer) and their initial score on a computer-test (taking into account their ability to work with the complex EHC-instrument).

*** Use of biomarkers and other physical measures in 3MC**

Joe Sakshaug, University of Michigan
Luzia M. Weiss, Max Planck Institute for Social Law and Social Policy, Munich

Health surveys have expanded their data collections over the last decades to include more objective measures of biological and physical health to supplement traditional self-reported health measures. Such objective measures, often referred to as biomarkers or physical measures, include the collection of biological materials (e.g., blood), anthropometric measures (e.g., height/weight), physical performance assessments (e.g., grip strength), and genetic data. Most recently, there has been an interest in the collection of these objective measures in a cross-national context. One particular example is the Survey of Health, Ageing and Retirement in Europe (SHARE) which collects biomarkers and/or physical measures in 20 countries in Europe (and Israel). Collecting biomarkers and physical measures cross-nationally poses many challenges in terms of achieving comparability across countries. For instance, logistical challenges range from different ethical requirements stated by the responsible ethics committees, different legal restrictions, to varying customs rules for shipping the equipment to the participating countries. With regards to data collection, comparability of the interviewer trainings in all countries has to be ensured, and varying consent rates over countries suggest a varying willingness to participate. Different legal and ethical requirements regarding the shipment, storage, and the analyses of biological samples can also arise. In this chapter, we expand on some of these challenges and highlight a number of important ways in which the use of biomarkers and physical measurements differ between national and cross-national surveys. We explore these issues in detail drawing heavily on the experiences of the SHARE as well as other cross-national studies, including the U.S. Health and Retirement Study, the English Longitudinal Study of Ageing, and the China Health and Retirement Longitudinal Study.

Thursday, July 28th, 11:00 a.m. - 12:30 p.m.

Session: Paradata in 3MC Surveys

Chair: Mengyao Hu, University of Michigan
Location: Michigan Ballroom I

Non-response analysis in the 6th European Working Conditions Survey

Aleksandra Wilczynska, Eurofound
Gijs van Houten, Eurofound
Mathijn Wilkens, Eurofound

With falling response rates in surveys, the non-response bias assessment and adjustment are of increasing importance for survey research. Recognizing the issue of non-response bias is especially relevant in the context of cross-national surveys, in which response rates may differ considerably from one country to another. Differences in the non-response bias between countries might jeopardise the comparability of the results. European Working Conditions Survey (EWCS) is a face-to-face survey of working conditions and quality of work and employment in Europe conducted each five years. In the 6th edition of EWCS conducted in 2015, the paradata were collected including not only the contact process information, but also interviewers' observations. The aim of this paper is to explore the non-response patterns in the 6th edition of the EWCS using these auxiliary variables, to establish similarities and differences between respondents and non-respondents. In the analysis we consider the continuum of resistance and classes models . The continuum of resistance model uses the group of respondents that required several contacts to be interviewed as a proxy for the group of non-respondents. The classes model makes an additional assumption about the heterogeneity of the non-respondents due to their accessibility and amenability.

Hence, it approximates the group of non-respondents that were impossible to contact with hard-to-contact respondents and the group of non-respondents that refused to be interviewed with the respondents who initially refused the interview. Using the interviewers' observations we test the assumptions of the models. Subsequently we apply the models to examine significant differences in outcomes of respondents with different response propensity.

Using paradata to monitor interviewers' instrument navigation behavior and inform instrument technical design: A case study from a national household survey in Ghana

Yu-chieh (Jay) Lin, University of Michigan
Kyle Kwaiser, University of Michigan
Gina-Qian Cheung, University of Michigan
Jennifer Kelley, University of Michigan

Many computer-assisted personal interview (CAPI) software captures paradata (i.e., empirical measurements about the process of creating survey data themselves), computer user actions, including times spent on questions and in sections of a survey (i.e., timestamps) and interviewer or respondent actions while proceeding through a survey. The paradata file contains a record of keystrokes and function keys pressed, as well as mouse actions. These paradata files are transmitted along with the survey data and are used extensively for quality assurance checks and reporting, particularly when the interviews are not recorded. This presentation uses data from the Ghana Socioeconomic Panel Study collaborated by Economic Growth Center at Yale University, the Institute for Statistical, Social and Economic Research at University of Ghana, and the Survey Research Center at University of Michigan. The study utilizes unique team management and travel structure, and has a complex instrument design. In addition, interviewers are allowed to interview respondents within the same sample unit without any particular order and to switch among varied interviewing components (i.e., household, personal, plot, enterprise sections) in a flexible fashion. Paradata is heavily relied on to monitor interviewers' behaviors. This presentation focuses on using keystroke data to monitor interviewers' instrument navigation behavior. We first reconstruct and categorize interviewer navigation patterns such as mid-section break-offs through varied interviewing components. These navigation patterns are then inspected for predictive power against data quality indicators such as response changes and non-response. Then, we analyze interviewer, household, and geographic characteristics and identify interviewer quality control metrics (e.g., interview length) to determine if interviewer behaviors and interview efficiency can be predicted by interviewer's overall team behavior or household characteristics, among all other information available. Finally, we will present how these analyses of paradata can be practically applied to improve interview efficiency and data quality of interviewer administered surveys.

Thursday, July 28th, 11:00 a.m. - 12:30 p.m.
Session: Quality Assurance and Quality Control 1

Chair: Agnes Parent-Thirion and Greet Vermeylen, Eurofound
Location: Great Lakes D

Too much or not enough? Assessing quality with weighting data in cross-sectional surveys

Frederic Gonthier, Sciences Po Grenoble, University Grenoble Alpes, France

The issue of quality improvement is usually addressed by scrutinizing the very first steps of the survey life cycle, such as coverage, sampling and non-response errors. Post-survey adjustements are less used, certainly because they are seen as a last resort in the error correcting process. Yet they can provide critical insights for researchers and survey practitioners engaged in quality assessment. Our paper is based on the European Values Study, a cross-sectional survey research program conducted every nine years since 1981. We use a cumulative dataset, including weights correcting for both gender and age at the national level. To embrace a coherent scope of countries, we select the twenty-six countries that participated in the last three waves. Similar sampling guidelines and weighting procedures permit robust comparisons between these countries. Has the quality of the EVS data increased from 1990 to 2008?

Contrary to under-representation, over-representation is not necessarily a sure hint of some degraded data. To assess separately over- and under-representation, we basically split the file in two separate sub-samples. Then we perform different multilevel models regressing on the two variables. Multilevel modeling is very relevant to examine how weighting diverges between countries and within different groups. Our analysis proceeds in four stages. First we use the so-called empty model to investigate differences within and across countries. The extent to which the magnitude of the weights varies from on country to the other gives useful indications to compare the quality of the national sampling processes. Next we regress on the over- and the under-represented respondents with gender and age, to compare how they account for the quality of the data, but also to assess their differential effects on the EVS countries. Our third step is to examine the influence of time. Since the regressions are performed on a pooled file, we add a dummy variable for each EVS wave to control for change and for the impact of the survey context over time. This should indicate in which country quality has, if so, increased the most. We further tackle data quality using macro-level variables. We supplement the file with variables from the EVS technical reports, such as response rate, number of languages fielded, mode of data collection, net sample size, number of visits. This innovative use of multilevel modeling allows us to finally track for the methodological levers of sampling and fielding efficiency at the national level.

Compliance and usage in the Generations and Gender Programme

Tom Emery, Netherlands Interdisciplinary Demographic Institute (NIDI)
Arianna Caporali, The French Institute for Demographic Studies (INED)

Launched in 2000 by the UNECE, the Generations and Gender Programme (GGP) is a longitudinal comparative survey of 18-79 years old in 19 countries in Europe and beyond run by a consortium of research institutions. It is based on a relatively decentralized management model and relies on considerable post hoc harmonization of data. The international core questionnaire is either adapted to the different national contexts or partly incorporated into existing surveys. Using data from the surveys administration, we examine the quality of compliance and standardisation in the GGP and whether this affects data usage. Firstly, we examine compliance by analysing the extent to which instruments from the core questionnaire were fielded within each of the 19 countries in the GGP. The results show that on average across countries, 66% of instruments in the core questionnaire were captured. Secondly, to examine usage, we take administrative data from the GGP website to capture the number of times each country dataset is downloaded. We supplement this with an analysis of the GGP bibliography and examine the number of times a country dataset is used in peer reviewed comparative publications (about 530 references). Finally, OLS regression analysis are presented to provide an overview of the association between compliance and usage, controlling for a number of contextual variables (e.g. number of IUSSP members, population with ISCED 8 in each country). The paper concludes with recommendations for future data collection activities and with reflections on the usefulness of analysis of compliance and usage in having an overview of the quality of comparative projects.

The European Social Survey's expanded framework for quality assessment

Katrijn Denies, University of Leuven (KU Leuven)
Koen Beullens, University of Leuven (KU Leuven)
Geert Loosveldt, University of Leuven (KU Leuven)

Since its inception in 2002, the European Social Survey has measured the attitudes, beliefs and behaviour patterns in over thirty nations, always prioritizing measurement equivalence. This goes hand-in-hand with an important place for quality assessment, i.e. in finding out which aspects of the survey life cycle could receive more attention in order to keep meeting the goal of being a survey of outstanding quality. By round 6, the ESS data and paradata were being used to draw up extensive quality reports about various aspects of fieldwork and its outcomes. For round 7 (data released late 2015/ early 2016), the round 6 framework for quality assessment has been expanded considerably and subjected to empirical assessment. In line with the Total Survey Error framework, the survey data are still scrutinized to reveal possible sources of differences between true population statistics and the obtained values. In addition, following the Total Quality Management paradigm, the entire process leading up to the final dataset is now also

being evaluated. This means that two main dimensions - representativity and measurement - are evaluated systematically from two perspectives: that of the process, and that of the output. The ESS quality assessment can therefore be seen as a thorough exercise in taking a broad view on the survey lifecycle and enriching it with cross-national and cross-rounds comparative components. This presentation will give a brief overview of the elements that are part of the new and expanded ESS quality assessment. For instance, as a process factor at the measurement level, the adherence to each step in the TRAPD (translation) procedure is mapped across countries, and as an output factor at the representativity level, the quality report looks into nonresponse per country. The core part of the presentation will focus on the most striking findings. As a first example, interviewer effects that vary strongly across countries are pointed out as the reason why ESS will pilot a survey among interviewers in round 8. Secondly, comprehensive graphs reveal considerable between-country differences in the length, intensity and planning of fieldwork. A last example that can be given here is the evolution of response rates across rounds in relation to other characteristics of fieldwork: only few countries achieve the target response rate, in spite of more refusal conversion, longer fieldwork periods, and more attempts per case.

Thursday, July 28th, 11:00 a.m. - 12:30 p.m.
Session: Questionnaire Design and Testing 2

Chair: Steve Dept, cApStAn
Location: Huron

* Cognitive interviewing methodology to examine survey question comparability

Kristen Miller, National Center for Health Statistics

Over the past decade, the need to assess cross-cultural comparability of questionnaires has generated increasing attention (Harkness et al, 2010; Fitzgerald, et. al, 2009). Do questions mean the same to all socio-cultural and linguistic groups represented in a survey? Are data elicited from questions capturing the same phenomena across all groups of respondents? While comparability of questionnaires has focused on the importance of translation and methods for producing accurate translations, a growing literature suggests that, outside of translated materials, respondents' life experience as well as socio-economic and cultural context can impact comparability (Miller et al, 2014; Willis, 2014; Miller and Willis, forthcoming). For example, respondents with little access to adequate health care, whether they live in a developing country or in a poor rural area of the United States, may be less able to accurately report having chronic conditions such as diabetes, high cholesterol, high blood pressure or emphysema. Respondents with strong religious beliefs may react more strongly or take offense to particular survey questions than those without such beliefs. The differences in the way respondents approach and process questions can affect data quality and undermine a survey's ability to accurately portray the population and the various subgroups within that population.

Issues of comparability necessarily pertain to validity and the consistency of validity across groups of respondents. Comparability is relevant to not only multinational surveys, but to any survey covering diverse populations. Focusing on cognitive interviewing methodology, this paper illustrates how analysis of in-depth, cognitive interviews can investigate the comparability of survey questions. Using examples of both international and national cognitive interview studies, this paper first describes the analytic goal of multi-cultural cognitive interviewing studies and then discusses relevant aspects of a cognitive interviewing study, including data collection (e.g. the structure of the interview and data quality), analytic techniques, strategies and tools for conducting such studies.

Traditionally cognitive interviewing is used as a pretest method to identify problems in questions prior to fielding a full survey. More recently, however, its use has expanded to include the study of construct validity (Miller, 2013; Miller et al, 2014). As a qualitative method that examines the processes and considerations used by respondents as they form answers to survey questions, the method identifies how those processes manifest within the various contexts of respondents' lives, experiences and perceptions. With a specific analysis, cognitive interviewing studies

can identify the basis of respondents' answers, indicating the phenomena or sets of phenomena that a variable would measure. With this analysis, comparability can then be examined by determining whether questions consistently capture the same phenomena across groups of respondents. The findings can be used to 'fix' questions before they are fielded, but they can also be used to inform the interpretation of survey data.

In this paper, examples from both national and international studies are used to illustrate:

1) how respondents from different socio-economic, cultural and lingual backgrounds can interpret questions differently,
2) how respondents with different personal experiences may interpret questions differently,
3) how cognitive interviews can be analyzed to determine the phenomena respondents include in their answer, that is, the various phenomena captured by survey questions (assessment of construct validity),
4) how cognitive interviews can be analyzed to determine whether various groups of respondents process questions differently (assessment of comparability).

In summary, this paper focuses on the qualitative method of cognitive interviewing as it pertains to question comparability. The paper illustrates the relationship between socio-cultural phenomena and survey data quality, specifically, how the socio-cultural and linguistic backgrounds of respondents can affect the question response process, which in turn can impact comparability. The paper will also illustrate how cognitive interviewing methodology can investigate comparability and provide useful information regarding question alteration and data usage.

* Sensitive questions in comparative surveys

Anna Andreenkova, CESSI-Russia (Institute for Comparative Social Research)
Debra Javeline, University of Notre Dame

The issue of sensitive questions is well-known and widely discussed in the methodological literature (T. Johnson 2002, Tourangeau and Yan 2007). Sensitive questions can increase the total measurement error of a survey (item non-response, misreporting or dishonest answers, or unit non-response) and may also have an indirect influence on subsequent questions and responses. In a comparative context, the differential impact of sensitive survey questions may undermine the comparability of survey data, bias results, and ultimately lead to misinterpretations of substantive conclusions (Javeline 1999).

We define topics and questions as sensitive if many respondents experience systematic emotional and cognitive difficulties due to the perceived pressure of social or cultural norms or legal or moral requirements. Sensitive questions differ from "threatening questions" - questions for which a particular answer could lead to political, social, economic, or moral sanction - and "questions with social desirability effect" - for which a particular answer is supported by social or cultural norms and can portray the respondent favorably to others and preserve his or her self-image. A tendency to offer socially desirable responses is often thought to reflect an individual's response style or personality (T. Johnson 2010) rather than the sensitive nature of the survey question.

The issue of sensitive questions in comparative context has not been sufficiently explored. Here we suggest an approach to studying cross-national variation in question sensitivity. We argue that methodological research on the cross-national and cross-cultural sensitivity of survey questions prior to the questionnaire design process can substantially improve data quality, decrease measurement error, and increase data comparability. This approach was tested in post-Soviet countries. The study includes three-stages. First, a list of potentially sensitive topics in each surveyed country was generated by a group of experts from different fields, and nonresponse in available cross-national surveys was analyzed. Seven broad categories of sensitive topics were revealed: family issues, political issues, financial and material issues, risk behavior, health, values (including religious and ethnic values), and knowledge. Second, respondents in national surveys based on random probability samples were asked to rate their level of ease/difficulty in answering questions on each issue. Third, cognitive interviews with individuals of different

social groups were conducted to understand the reasons for the sensitivity of different questions, the related emotions, and the process of response formation for different types of sensitive questions.

The data show that three out of seven issues evaluated in the survey appeared to be similarly sensitive in all countries: finance and ownership, risk or deviant behavior (binge drinking and smoking), and health issues. Four other topics demonstrated country-specific sensitivity: political behavior and political views, family structure, knowledge, and values (general and religious). We hypothesize that cross-national variations in sensitivity are related to the level of religiosity of the country, level of urbanization, degree of gender equality and distance in perceived gender roles (masculine vs. more equal societies), and type of political regime. Based on the results of qualitative interviews, we propose that survey questions may be sensitive due to the perceived violation of a) judicial or political laws, norms, or requirements; b) social interaction or moral norms; or c) cultural norms of communication and the perception of "privacy". When the level of sensitivity and the type of sensitivity is similar for different countries, the sensitivity leads to measurement error in all countries, but it does not have a strong effect on data comparability. If sensitivity is different in different countries, the impact on cross-national comparisons can be large and needs to be taken into account in analysis. Understanding the reasons for sensitivity and whether they vary cross-nationally or cross-culturally can help to assess whether sensitivity has any influence on data comparability and to illuminate ways to address this problem.

Framing numeric questions using evidence from observational research
Meghann Jones, Ipsos

Since the goal of international development programs is typically to improve the financial situation of beneficiaries, studies in the field frequently focus on, or include a component of, financial information such as household income or business revenue. Unfortunately, we frequently observe that the quality of the numeric data collected is poor. There are several reasons for this , numeracy may be a problem for program beneficiaries (and even interviewers), where records are not kept recall is problematic, data collected by PAPI can be unclear to data entry teams, and data collected using CAPI might get bad numbers from an accidental extra digit. However, one of the most pressing problems we observe is that the question itself does not make any sense to the respondent: perhaps we are asking about household income with the option for only monetary responses, when the respondent relies mainly on exchange to procure good for his family; perhaps we are asking the respondent to report on business profit, when she thinks only about cashflow. This is particularly problematic for multi-country studies where there is a desire for consistency across markets, but context, language, and cultural differences mean than questions are interpreted differently. In this paper we will demonstrate the complex nature of applying financial questions to multiple contexts, and discuss how qualitative and observational research can be used to design questions that enable context to be taken into consideration and accurate data be collected.

Thursday, July 28th, 11:00 a.m. - 12:30 p.m.
Session: Response Styles
Chair: Anna Andreenkova, CESSI-Russia (Institute for Comparative Social Research)
Location: Great Lakes E

*** Cross-cultural comparability of response patterns of subjective probability questions**
Sunghee Lee, University of Michigan
Florian Keusch, University of Mannheim
Norbert Schwarz, University of Michigan
Mingnan Liu, Survey Monkey
Z. Tuba Suzer-Gurtekin, University of Michigan

Questions soliciting respondents to estimate chances for future events to happen (e.g., inflation, job loss, mortality, product purchase) have become increasingly popular in surveys. The expectation question usually uses the following wording within [FUTURE TIME FRAME], what are the chances that you will [FUTURE EVENT]? Respondents are asked to rate their expectation for a certain event happening on a numeric scale of 0 to 100 indicating probabilities. Answers to these questions have been shown to predict actual behaviors, the main driver for the increasing popularity. Apart from the debate on whether or not people can carry out probabilistic reasoning, the known measurement error on expectation questions is response heaping at a focal point, typically 50. However, these findings do not consider two important yet implicit premises related to cognition germane to answering expectation questions. First, expectation questions assume that respondents organize their cognition relevant for the future. The second assumption is that the way in which the cognition is organized reflects realistic outlooks. While these assumptions themselves may raise concerns of measurement error, they may raise even larger concerns for cross-cultural surveys because of cultural differences in cognitive processes. In particular, the following two cultural dimensions matter the most for expectation questions. The first dimension is time perspectives that influence cognition organization "whether people process their personal experiences in their memory in relations to past, present or future. This means that respondents with past- or present-oriented cultural backgrounds may find the expectation questions more difficult than those from future-oriented cultures. The second dimension is the locus of control that differs between individualistic and collectivistic cultures. Typically, individualistic cultures are found to score higher on primary and direct control, whereas secondary control, which provides individuals with feelings of control by aligning with more authoritative figures, is more prevalent in collectivistic cultures. As the locus of control is internal in individualistic cultures, it is likely that those from individualistic cultures have more realistic and controlled views about their own future than those in collectivistic culture whose locus of control is external. In the measurement of expectation questions, these cultural orientations are likely to produce two distinctive response patterns. Differences in time perspectives are equated with differential cognitive difficulties from respondents perspectives when answering expectation questions, resulting in differential item nonresponse rates, where future-oriented cultures are associated with lower item nonresponse rates. Respondents with a weak sense of control may choose deterministic (i.e., unrealistic) response points, such as 0 or 100. On the other hand, those with a strong sense of primary control are likely to have more realistic views about the expectations, making their responses to expectation questions distributed away from the deterministic answers of 0 or 100 and potentially following a type of normal distribution. This tendency with the locus of control is likely to be manifested through different response heaping points. Further, the cultural distinctiveness in response patterns of expectation questions is likely to be more apparent for future events that respondents have less personal data for (e.g., retiring in five years year vs. stock price going up in five years). In sum, expectation questions are not free from measurement error, and especially in cross-cultural contexts, they are subject to noncomparabilty of measurement error. Moreover, noncomparability is larger for future events that are less relevant to respondents. This study attempts to provide theoretical backgrounds on these cultural orientations pertinent to response patterns of expectation questions by introducing the cross-cultural cognitive psychology literature. We will then empirically examine the influence of the cultural orientations on expectation questions in different topics ranging from inflation to mortality using various data sources, such as the Surveys of Consumers (SCA), the Survey of Consumer Finances (SCF), the Health and Retirement Study (HRS), the English Longitudinal Study of Ageing (ELSA), and the Survey of Health, Ageing, and Retirement in Europe (SHARE). SCA includes a wide range of expectation questions related to economic behavior and covers culturally distinctive groups, although it is conducted in one country. SCF includes questions on expectations of moving. The last three surveys are methodologically comparable and address multiple countries with varying cultural orientations. They include not only various expectation questions but also personality measures that operationalize cultural orientations of time perspectives and sense of control. This allows us to examine the effect of cultural orientations at the group (e.g., country, language, race and ethnicity) level as well as the individual level without potential confounding effects due to methodological noncomparability. Given the data availability, the cultural orientations described above lead to the following three hypotheses: 1) Item nonresponse rates on expectation questions are lower for future-oriented cultural groups (e.g., White Americans) than past- or present-oriented groups (e.g., Hispanic Americans and Black Americans); 2) Responses to expectation questions heap at 50 for cultural groups with strong sense of direct control (e.g., Germanic language speaking groups), while it is 0 or 100 for those with weaker sense of control (e.g., Romance

language speaking groups); and 3) Culture-specific item nonresponse and response heaping patterns are smaller for expectation questions that ask about events for which respondents have concrete and relevant personal data than insufficient and ambiguous data. These hypotheses will be examined using bivariate analysis as well as multivariate analysis. A multi-level analysis will also be considered to account for interviewer-level characteristics and country- or cultural-group level effects. Particularly for mortality expectation questions in HRS, ELSA and SHARE, we will also examine the association between response patterns and actual mortality in subsequent time points as some of the data listed above include individual level data on time orientation and sense of control which can serve as a longitudinal component applicable for this analysis.

* Lost in translation? An empirical study on the impact of translation on respondents' use of response scales and response distributions

Ting Yan, Westat
Mengyao Hu, University of Michigan

Comparison of data from surveys conducted in the multinational, multiregional, and multicultural (3M) contexts is challenging. Any observed differences in answers may reflect the actual differences in behaviors and attitudes, differences in survey response process due to culture and/or language, or an unknown mix of both. Translation is both necessary and critical when conducting surveys in the 3M contexts. However, it does introduce an additional source of measurement error to 3M surveys. Translation of survey questions and response scales may affect how potential respondents understand the questions and the scales, what information they retrieve, how they put retrieved information together to generate an estimate or judgment, and how they map their estimate and judgment to one of the response categories. In particular, translation of response scales may affect how respondents use the scales to generate and/or map an answer. As a result, translation could confound any observed differences across studies conducted in different languages. Unfortunately, literature on the evaluation and assessment of translation is limited. Quantitative evaluation and assessment of translation is even more limited. This paper attempts to fill the gap by conducting an empirical study examining the impact of translation of response scales on how respondents use the scale as well as the resultant response distributions. This paper studies the translation of two response scales used to measure self-reported health (SRH). 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. SRH uses either a bipolar scale ranging from very good, good, •fair, poor •to very poor •or an unbalanced unipolar scale ranging from excellent, very good, good, fair, •to poor. SRH is included in four different surveys administered in Chinese. Two of the surveys, the China Health and Retirement Longitudinal Study (CHARLS) and the Taiwan Social Change Survey (TSCS), include an experiment that randomly assigns one half of the respondents to the bipolar scale and the other half to the unbalanced unipolar scale. The third study, The Chinese Family Panel Survey (CFPS), uses the unbalanced unipolar scale whereas the Chinese General Social Survey (CGSS) uses the bipolar scale. The two response scales are translated very differently among the four studies. Drawing on data from these four Chinese surveys, we will examine how different translations of the scales affect respondents' use of the scales and the resultant answers. Preliminary results indicate that different translations of the two scales did produce different distributions of self-reported health. In addition, we will also compare data from CHARLS to its sister survey, the Health and Retirement Study (HRS), conducted in the U.S. to further explore the role of translation in observed differences between CHARLS and HRS. We will discuss the impact of translation on respondents' understanding of the response categories and their choice of response strategies 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 conducting comparative studies employing surveys conducted in the multinational, multiregional, and multicultural (3M) contexts.

Thursday, July 28th, 2:00 p.m. - 3:30 p.m.
Session: Analysis Methods and Tools 3

Chair: Timothy P. Johnson, University of Illinois at Chicago
Location: Michigan Ballroom II

The UNECE international statistical framework for measuring quality of employment

Christian Wingerter, European Commission, Eurostat

A United Nations Economic Commission for Europe (UNECE) expert group consisting of international organisations like the International Labour Organisation and Eurostat and many national statistical offices recently has developed an internationally agreed framework on measuring quality of employment. It offers a coherent structure for measuring employment quality and provides practical guidance for compiling and interpreting a set of almost 70 indicators comprehensively measuring employment quality. It also facilitates the assessment of employment quality from an international comparable perspective. The framework is based on a theoretical concept defining seven dimensions in order to map employment quality in all its aspects and is setup to be adequate for various labour market conditions worldwide. The presentation will explain the underlying rationale of the framework, the theoretical concept with its seven deducted dimensions and introduce some indicators in detail. For the latter illustrative data will be presented with reference to the European Union member states. There will also be a report about a first endeavour to make data on quality of employment accessible on an international level by the online database of Eurostat.

Why do people trust political institutions in the presence of corruption?

Irena Schneider, King's College London

It is often theorized that having to use bribes or social networks to obtain basic public services diminishes a person's trust in political institutions. This appears not to be the case in rarely studied portions of the post-Soviet space, raising interesting questions about how and why governments retain public support in dysfunctional institutional conditions. I investigate the relationship between political trust and corruption in a sample of 35 former Soviet and European countries using the 2010 Life in Transition Survey II conducted by the World Bank and European Bank for Reconstruction and Development. Building on a growing political economy literature on clientelism, I hypothesize that people will be less dissatisfied with political institutions in the presence of corruption when economic times are good. I run a multilevel structural equation model to assess the extent to which country-level and regional economic conditions moderate the effect of corruption perceptions on political trust in the sampled territory. To establish measurement equivalence among diverse cultures and regime types, I use sets of metrically invariant latent variables across regions. This study offers a novel test of the theory in authoritarian country samples, and is among the first attempts in the literature to simultaneously account for measurement error, equivalence and the hierarchical nature of the data.

Quantile regression as a tool for cross-national and comparative survey research

Robert A. Petrin, Ipsos Public Affairs
Joseph Zappa, Ipsos Public Affairs
Meghana Raja, Ipsos Public Affairs

Survey research often involves making comparisons across multinational, multiregional, and multicultural contexts. When these contexts are qualitatively different in terms of social, cultural, economic or other factors, standard methodologies for analyzing survey data and making cross-context comparisons of dependent variables can be inadequate. The reason is because many conventional procedures for analyzing survey data only provide a way of evaluating differences in means across analytic groups, and in doing so do not account for substantial qualitative variation in the distribution of outcomes across these groups. For example, in the case of quantitative survey

research aimed at understanding acceptable pricing of public good or service, it is not just the mean level of ideal prices or perceptions of the value of a program or service that varies across countries or regions, but the distribution of these measures as well. Substantial variation in the dependent variable across contexts often requires researchers applying conventional survey data analysis methods to ignore contextual differences altogether. The result is a loss of the richness inherent to multinational survey responses, as well as inaccurate inferences and potentially misleading understandings of differences across these contexts. These challenges are exacerbated when survey research layers a longitudinal component on top of a multinational or multiregional design. Accordingly, this paper presents techniques based on quantile regression (QR; e.g., Davino et al., 2015; Koenker, 2005; McMillan, 2012) which can remedy some of the challenges which arise in multinational and multiregional comparative research. The paper begins with a series of conceptual illustrations and simulations to highlight when the assumptions upon which conventional statistical methods for analyzing multinational and multiregional survey data break down. It then outlines the basics of QR in the comparative research context, relating QR to conventional parametric and semiparametric regression techniques for understanding and adjusting for contextual differences when analyzing survey data (e.g., Ruppert et al, 2003). Then, using original empirical data from a global pricing study, as well as data from a multinational longitudinal evaluation of a women's entrepreneurship and business skills program, we present how QR can be adapted to allow meaningful comparisons across countries and regions which reflect the particulars of actors in those markets, as well as program roll-out. Specific attention is placed on Bayesian implementations of QR, which allows population-valid estimates to be obtained from common types of sample survey designs (e.g., Gelman et al., 2013; Si, 2012).

Survey quality prediction: A tool to quantify measurement quality
Willem Saris, Universitat Pompeu Fabra
Daniel Oberski, Tillburg University
Melnie Revilla, Universitat Pompeu Fabra
Diana Zavala-Rojas, Universitat Pompeu Fabra

Quality standards in survey research suggest that the Total Survey Error (TSE) should be minimized. In 3MC survey research, this challenge is magnified because quality standards should be high and homogeneous in all settings where the survey project is developed. However, it is difficult to quantify "quality" because many decisions taken when the survey is designed, such as the question format, the mode of data collection, translation decisions, among others have an impact in data quality. This paper presents the approach behind the Survey Quality Prediction (SQP) software, a semi-automatic program that allows predicting the measurement quality of survey questions in more than thirty countries. Measurement quality is defined in SQP by the product of the reliability and the validity of a survey question. Prediction of the measurement quality is based on a meta-analysis over 3,000 multi-trait multi-method experiments (MTMM) in more than twenty languages and a coding system of over 70 specific characteristics of survey questions. SQP aims to fill a gap in methodological research by being a tool that quantifies the quality of a question in a survey. As it is possible to obtain a prediction for a same question in different languages, SQP predictions are a powerful source of evidence that allow cost-effectively to obtain systematic and statistically relevant information about the questions to correct for measurement error in comparative survey research.

Thursday, July 28th, 2:00 p.m. - 3:30 p.m.
Session: Language
Chair: Rachel Caspar, RTI International
Location: Michigan Ballroom I

*** Language to interview—Comparability in multi-lingual context**
Anna Andreenkova, CESSI (Institute for Comparative Social Research)

Language of an interview is not always the obvious choice in many cultures, nations and political contexts although this issue is often ignored in single-country and comparative surveys. Researchers are confronted with a choice of language for an interview in regions with multi-lingual population. And the size of such population is growing due to globalization of education, migration, cross-ethnic marriages and other factors. For example, in the countries of the former Soviet Union, the share of multilingual population defined conservatively as a population with more than one native language ranges from the lowest of 7% in Russia to 73% in Belarus; the share of such population is over 50% in Ukraine, Kazakhstan, Kyrgyzstan, Moldova and some other countries. So the situation when the choice of language for an interview is required is very frequent and not equal among countries. Thus, for comparative surveys it would be very important to have a clear rationale and a basis for making this choice. The decision about the interview language is taken in two stages of the research project - on the planning or design stage and on the field stage. On the design stage researchers make a decision regarding the languages a survey instrument needs to be translated into. The analysis of major comparative and large-scale national surveys shows that the choice of language to translate is either done without clearly explained and documented rationale or based on assumptions about major • languages in each country. For example, in ESS the translation is done into languages considered as main • or native by min 5% of population of each country. On the design stage of research project a lot of other types of considerations can influence the choice of languages for interviews - linguistic, organizational (cost of translation, feasibility of high quality translations, availability of interviewers with specific language skills, etc.), social and political. On the field stage in most surveys that we analyzed, the choice of language is assumed to be done by respondent. But this choice depends on the set of languages pre-defined by researcher, cultural norms and traditions in interaction, functional image of the language among a target group, linguistic skills of actors, type of interaction between an interviewer and a respondent, direct influence of an interviewer on a respondent, or based on personal preferences. Possible consequences of the wrong choice of an interview language or the use of suboptimal criteria for choosing a language are, firstly, an increase in the sample error: a/ the under-coverage issue (exclusion of some groups of population due to the language barrier), and b/ the decrease of respondents cooperation; and, secondly, an increase in the measurement error: a/ lower reliability of information if the language skills of a respondent or an interviewer are not sufficient or b/ systematical bias of information from respondent's side if a chosen language implies high social pressure characteristic or is not politically and socially neutral in a given environment. For the decision which languages to use in a survey and which translations to be done on the survey design stage, we argue that we need to collect and to use the detailed information about the functioning of different languages among surveyed groups and types of linguistic skills required to complete the survey. A number of survey experiments which we conducted in the countries of the former Soviet Union showed that multi-lingual respondents often have different function and skill for different languages. Multi-lingual respondents often use different languages for different tasks - for inter-family communication, for outside communication, professional communication, media consumption etc. As a result respondents form the opinion which language is more appropriate for use in different social situations and also learn different vocabulary in different languages. Not only respondents use languages differently, they often may have and use different linguistic skills for different languages "either oral or writing or reading. We argue that all this information should be taken into account when taking decision which language to choose for particular type of survey (self-completion, face-to-face interview, telephone interview etc.) and also for particular subject of the survey which leads to specific vocabulary requirements (everyday words, political vocabulary, religion, job, etc.). Regarding the last stage of language selection – respondent's preference, we also argue that currently respondents make the choice without prior information on language requirements for the interview. Respondents are not informed which linguistic skills will be required and which vocabulary will be used. Our experiments show that in about one forth of multi-lingual cases respondents made non-optimal choice of the interview language compared to the choice that could have been done based on assessment information on language usage and skills for given respondent. In majority of cases the wrong choice led to serious problems during the interview: attempts to switch the language during the interview, use of mixed language - combination of different languages, higher Do not know • responses and, in some cases, higher response style • effect, i.e. less differentiated answers on response scale. We argue that both researchers and respondents should make more informed decisions on the choice of interview language based on the information about the function of different languages in respondents life and linguistic skills required for

particular type of survey, information collected prior to the survey among the multi-lingual group level and also on the individual level prior to beginning of an interview.

* Can the language of a survey interview influence respondent answers?

Emilia Peytcheva, RTI International

Surveys in many countries have to allow administration in multiple languages in order to avoid excluding large parts of the population. The U.S. Census Bureau projects that between 2020-2050, the Hispanic-origin population will contribute to 62% of the population growth, reaching 25% of the total population in the U.S. by the year 2050 (Day 1996). Currently, many national surveys offer respondents a choice of language (e.g., the New Immigrant Survey; the National Survey of Latinos; the National Latino and Asian American Study). However, the use of different languages to measure the same phenomena inherently adds a potential source for differences between those interviewed in English and those interviewed in another language. The existing cross-cultural measurement literature focuses on effective translation to achieve equivalent survey instruments (e.g., Harkness, Van de Vijver, and Mohler 2003), but has overlooked the role of language as a factor affecting the response formation process and cueing related cultural values and norms. Recent psychological and linguistic research suggests that language affects respondent's reference frame, potentially influencing how respondents perceive the intent of the survey questions and their affective characteristics, including sensitivity and need for socially acceptable answers. For survey practitioners this would suggest that bilingual bicultural respondents may answer the same question differently, depending on the language of interview. Indeed, we know from carefully controlled laboratory studies that language can have such influences on tasks that are designed to assess them (for example, Trafimow, Silverman, Fan and Law,1997; Schrauf and Rubin, 1998; Schrauf and Rubin, 2000; Marian and Neisser, 2000; Ross, Xun and Wilson, 2002). However, the psychological research provides little insight into whether these influences are of practical relevance to survey research; conversely, the findings in the survey methodology literature fail to isolate the influence of language per se. This paper attempts to fill in the existing gap by presenting an empirical estimation of language effects in two national surveys of immigrants. We use data from the National Latino and Asian American Survey (NLAAS) in which a sample of Spanish-English bilingual respondents was randomly assigned to English or Spanish administration of the survey, allowing direct estimation of the effect of language, and data from the New Immigrant Survey (NIS), where statistical methods such as propensity score modeling allow us to estimate the effect of language when respondents self-select themselves into a language of interview. We examine the hypothesis that language of survey administration will affect responses among Hispanic bicultural-bilingual respondents in the U.S. if and only if the two cultures differ in their social desirability norms relevant to the questions being answered. Two sets of items are of interest across the two surveys " questions on highly stigmatized topics in the Hispanic culture, but not the American culture (such as, mental health and alcohol use), and demographic questions, where no language effects are expected. We found evidence that bilingual respondents provide different responses depending on the language of interview for questions that prime culture-related norms and, as expected, no such effects for questions that are related to demographic characteristics, such as marital status, or living situation. Specifically, in the NIS data we found significantly lower reports of alcohol consumption for Hispanic bilinguals interviewed in Spanish compared to those interviewed in English. Interestingly, we also found significant differences in the number of biological children reported in the two languages " consistent with the central value of familism in the Hispanic culture, we found significantly higher reports when respondents were interviewed in Spanish. Similar results were obtained from the NLAAS, where respondents interviewed in Spanish, endorsed more the value of familism, resulting in significantly different scores on a family pride scale. Similarly, the reported age at first alcohol consumption was higher for those interviewed in Spanish (consistent with more conservative alcohol norms and attitudes among Hispanics relative to Whites, see Cateano and Clark, 1999), but failed to reach statistical significance. The implication of such results for current national surveys that sample ethnic minorities and immigrants is that language assignment should be informed by the goals of the survey questions and leaving the choice of language to a bilingual bicultural respondent may affect data quality. Ideally, researchers would be able to inform language assignment based on knowledge about domains where cultural differences and the direction of such differences may be expected, or depending on what

respondent cultural identity is of interest. If such knowledge is not available, random assignment of bilingual respondents to a language would at least allow estimation of language effects.

* Working towards parallel meaning of different language versions of survey instruments: Do monolingual and bilingual cognitive testing respondents help to uncover the same issues?

Patricia Goerman, U.S. Census Bureau
Mikelyn Meyers, U.S. Census Bureau
Mandy Sha, RTI International
Hyunjoo Park, RTI International
Alisú Schoua-Glusberg, Research Support Services

Many researchers follow the "rule of thumb" that translated survey instruments should be cognitively tested with only monolingual and limited source-language proficient respondents. One common rationale for this is that fully-bilingual respondents, or those dominant in the source language, might be more likely to understand or less likely to notice a poorly translated phrase, which might, for example, mimic source language word order conventions. They might also be more likely to understand inappropriate literal translations for concepts that do not existent in the target language. Assuming these concerns are legitimate leads many researchers to conclude that testing translated questionnaires with only fully-bilingual respondents could cause cognitive researchers to "miss" problems in the translation. The decision about which language proficiency levels are acceptable in cognitive testing of survey translations is important in particular for cost reasons. It takes more time and resources to recruit monolingual and less acculturated respondents. The main goal of this research was to explore what types of respondents can help the most with cognitive pretesting of a Spanish-language survey instrument in the U.S. We designed empirical research to examine whether the number and types of cognitive interview findings vary by language proficiency level of respondents. We tested a U.S. Census Bureau survey instrument in the interviewer-administered, CAPI mode by conducting cognitive interviews with 39 Spanish-speaking respondents, half of whom were monolingual and half of whom were bilingual Spanish and English-speakers. We found that bilingual respondents identified most of the same problems as monolingual respondents, but not necessarily with the same frequency. In addition, there was a small number of issues that were only problematic for either the monolingual or bilingual respondents. In these cases, testing with only one group or the other would have masked these issues. On the whole, including only bilinguals in interviewer-administered, CAPI, cognitive testing may help to reduce research costs and timing, but some issues may not be revealed. More research should be done to compare cognitive testing done with monolingual and bilingual respondents in additional languages to see if these findings are generalizable. It would also be interesting to compare the findings across different survey modes to see if monolingual or bilingual respondents are able to uncover problems with survey questions differently depending on mode.

How do language differences between interviewers and respondents affect data quality in Africa?

Charles Lau, RTI International

A premise of in-person surveys is that interviewers and respondents speak the same language. But the great linguistic diversity of many African countries may pose difficulties for interviewers and respondents to communicate with each other. In linguistically diverse countries, it may be infeasible to translate surveys into all possible languages. In some cases, respondents with limited fluency in a language may attempt to complete the survey. In other cases, interviewers may attempt to communicate in a language where they lack proficiency. Compounding this problems is the fact that interviewers are typically recruited from urban centers, and have limited proficiency in minority languages. This study investigates how linguistic differences between interviewers and respondents affect data quality. I analyze data from the 2008 Afrobarometer Surveys (n = 27,713), in-person, paper-and-pencil surveys about political attitudes conducted in 20 African countries. The Afrobarometer data contain three measures of language: (1) the native language of the interviewer; (2) the native language of the respondent; and (3) the language in which the interview was conducted. Using these data, I describe the correspondence between these three languages, and show how linguistic differences vary across countries and respondent demographics. The analysis also investigates the

associations between linguistic differences and several indicators of data quality, such as item non-response, inconsistent attitude data, and acquiescence.

Thursday, July 28th, 2:00 p.m. - 3:30 p.m.
Session: Quality Assurance and Quality Control 2

Chair: Agnes Parent-Thirion and Greet Vermeylen, Eurofound
Location: Great Lakes D

The application of a quality assurance strategy in the 6th European Working Conditions Survey: Experiences and reflections

Sally Widdop, Ipsos MORI
Gijs van Houten, Eurofound

Several methods for quality assessment and assurance in statistics have been developed in a European context and the importance of assuring the quality of survey data is also increasingly being recognised. As noted by Lyberg and Stukel in 2010, many cross-national surveys have tried to incorporate quality assessments into their activities - with mixed success. This paper explores the approach to quality assurance taken on the 6th wave of the European Working Conditions Survey (EWCS). This survey seeks to measure the working conditions of employees and self-employed in Europe and is funded and carried out by the European Foundation for the Improvement of Living and Working Conditions (Eurofound). Fieldwork for the 6th wave of the survey was conducted in 2015 by Ipsos who interviewed 43,000 workers in 35 countries. The quality assurance strategy adopted for the 6th EWCS was based on Eurofound's pre-existing quality assurance framework - consisting of quality assurance reports and quality assessments. It was specifically developed for the 2015 survey with input from Ipsos who elaborated on and provided detail to the quality assurance strategy proposed by Eurofound. The quality assurance strategy for the 6th EWCS utilised the quality criteria adopted by the European Statistical System , namely: relevance, accuracy, timeliness and punctuality, accessibility and clarity, coherence and comparability. It also incorporated elements of the cross-cultural survey guidelines and the Total Survey Error framework. In total, around 140 explicit targets were set, including 'hard' requirements as well as 'softer' ambitions. These were applicable to each of the key stages in the survey life cycle , e.g. sampling, questionnaire, translation, interviewer training, fieldwork, data entry & data processing, weighting and micro-data. In this paper we will discuss the underlying rationale used to identify and define the quality indicators in the quality assurance strategy as well as reflecting on the application of these in the 6th wave of EWCS. We will assess our how the approach might be useful for cross-cultural surveys in general and consider possible improvements.

FINCA ValiData - Using real-time algorithms to improve field data collection

Scott Graham, FINCA
Anahit Tevosyan, FINCA

Field interviewer error (and fraud) can result in deeply flawed data sets, putting the entire research enterprise at risk. These problems are especially acute when data collection takes place in remote locations, where field supervision is difficult. Once a survey is done, it may be impossible to address apparent errors, which can render part or even the whole data set non-returnable. A battery of statistical techniques are used to address missing values, outliers, and problematic data, but this normally happens long after fieldwork completion. The process of ex-post data cleaning can itself introduce errors in the dataset, while it drains an analysts' time and resources , particularly when dealing with large, multi-site surveys. To address this problem, we have developed an automated, real-time data management platform, ValiData, which performs a range of statistical functions in order to flag outliers and behavioral anomalies, and thereby trigger corrective action while the survey is still in the field. This approach eliminates the risk of having problematic and outlying responses, ensures validity of the data and keeps survey statistics at programmed levels. From the analysis perspective, it saves the time that would otherwise be spent on

manual statistical routines, and eliminates the systematic and random errors that are introduced in ex-post cleaning procedures. From the field management perspective, it allows survey managers to direct their attention to the sources of fabricated and problematic data as they are arising. It also alerts field staff to the fact that mistakes are being caught, instilling greater vigilance for data quality throughout the survey team. Alongside common statistical measures such as IQR and standard deviation from the mean, ValiData employs robust regressions, conditional probabilities and logistic modeling to detect both numeric and categorical data anomalies. It also detects behavioral anomalies by executing decision tree classifiers, machine-learning algorithms and advanced clustering techniques on both survey variables and meta-data. Using these techniques, ValiData is able to detect falsified responses, prompting researchers to take proactive steps to protect the integrity of the survey. In this presentation, FINCA will describe the technical implementation of ValiData through a case study of its use in large-scale surveys in Africa, MESA and Latin America. The case study will highlight the research management challenges that presented in the field, the implementation and role of automated data analytics and their impact on the data set. Time allowing, the presentation will also include a live demonstration of the platform.

The European Working Conditions Survey Series over time: Lessons over time
Agnes Parent-Thirion, Eurofound
Greet Vermeylen, Eurofound

The European working conditions survey (EWCS) produced by Eurofound has been measuring working conditions and job quality since 1991. Six editions have taken place: topic coverage as well as country coverage have extended at each edition. The last edition, 2015 covers 35 countries : procedures implemented to ensure cross country comparability will be presented. The presentation will discuss reasons for monitoring working conditions and their contribution to policy concerns, present the specificities of the EWCS and its consequences on survey design, examine the topical scope of the survey and compare it to so called 'validated ' questionnaires extensively used in epidemiology, reflect on the development of a survey module dedicated to job quality.

Thursday, July 28th, 2:00 p.m. - 3:30 p.m.
Session: Questionnaire Design and Testing 3
Chair: Steve Dept, cApStAn
Location: Huron

*** Multi-national study of questionnaire design**
Henning Silber, GESIS

Large-scale international survey projects based on probability samples have been rare in the past because of the high requirements with regard to funding and manpower. However, with the rise of nationally representative online samples, both financial and labor requirements have gone down dramatically. In the near future, we expect to see many international studies that are organized by small international teams. In this chapter, we discuss challenges and solutions for arising problems in such decentralized survey teams. This discussion is based on our experience with organizing a study that tested principles of questionnaire design in 14 countries around the world. The goal was to gauge country-specific differences in response behavior, satisficing, and social desirability response bias (DeMaio 1984; Johnson & van de Vijver 2003; Krosnick 1991; 1999; Lensvelt-Mulders 2008). These survey experiments were conducted in Canada, Denmark, France, Germany, Japan, Iceland, the Netherlands, Norway, Portugal, Sweden, Taiwan, Thailand, the United Kingdom and the United States. The data have been collected during the same time period from national random probability online samples. In detail, we will present our study design and study objectives, challenges during the recruitment process for collaborating organizations, the various sampling strategies, and challenges faced during the organization of project. The chapter will conclude with a section on each participating organization describing the country specific survey design. Study design: Each sample in this research was a random probability sample of the general population of all adults living in the particular country aged 18 and

older, with everyone in the country having a known, non-zero probability of being participating in the survey. Samples of subpopulations such as samples of students were not acceptable, nor were data collected from non-probability samples of people who volunteer to complete surveys for money and are not randomly sampled from the nation's population. However, since survey participation, especially online, usually decreases sharply in the oldest age cohorts, samples with an upper age bound of around 70 years were acceptable. In this section we will discuss the study design, the sample requirements and the analyzing plan. Translation and requirements: Translation from the questionnaire in English into a language or languages suitable for use in a country had to be done using the TRAPD procedures developed by Janet Harkness (see Harkness 2003; 2007; Harkness, Villar, and Edwards 2010; http://ccsg.isr.umich.edu/translation.cfm). In every country, at least two translators with a background in survey research separately drafted a full translation of the questions, and then the translators together with the national project head discussed their drafts to agree on the final translation. Additionally, we asked collaborators to send us the translation of each translation team member and screenshots of the programed questionnaire before we authorized them to begin the data collection in their country. This section will focus on the translation methods and evaluation tools used. We also will emphasis challenges the various countries experienced during the translation process. Study goals The primary study goal was to run a set of classic experiments (see Schuman & Presser 1981) for question evaluation. More precisely, we proposed to run eighteen split ballot design question experiments. The experiments tested for the differences in response behavior by altering (a) the order in which the response options and (b) the questions are presented. Moreover, we tested (c) for differences caused by acquiescence (a tendency to agree with any presented statement), (d) for effects of different no opinion filters (filter 1: not enough information; filter 2: no opinion; filter 3: don't know), (e) for differences in response behavior, if the questions are introduced by some/others (e. g., Some people think that it is healthy to eat chocolate every day, other people think that it is not healthy to eat chocolate every day. What is your opinion on this? •), (f) for question balance effects (balanced questions are completely neutral), and (g) for the impact of conversational conventions on response behavior (e.g., by starting the question with a negation). This section will outline the selection criteria of the experiments and we will, additionally, discuss various reviews from researchers in the different countries and how we adapted the experiments for the specific national context. Recruitment process: The project started out small and grew organically during three different stages. First, we submitted our study proposal to open calls in different countries (e. g, LISS, IP, Citizen Panel, and ELIPSS). Second, we personally asked various researchers whether they are interested in joining our project team to run the study in their country. Third, we released an open call for participation through the AAPOR and WAPOR network. During each stage of the project teams across many countries joined the project. A lot of researchers were interested in participating but because every organization has to provide its own funding, many organizations could not participate. This section will explain how our recruitment strategies varied over time. The section will also evaluate the different recruitment strategies and discuss alternatives routes. Project organization: The project is organized in two groups, the core project team and the global project team. The core project team has four members who developed the study idea and are responsible for recruiting new countries and the project management. The global project team includes every participating researcher and it will release joined publications. Today, it includes more than twenty researchers from fourteen countries. We will discuss challenges we had to overcome in coordinating both an international core team and the large project teams. This section will focus on solutions we found to increase the efficiency of such a large-scale international project. Participating countries: The recruitment process, financing, and study organization varies between countries. We will outline the approach taken by each collaborating institution and discuss challenges that various organizations had to face.

Cross-cultural instrument development and testing

Mousumi Sarkar, IMPAQ International, LLC
Alisú Schoua-Glusberg, Research Support Services
Maurice Kugler, IMPAQ International, LLC
Jean Paul Petraud, IMPAQ International, LLC
Rocco Russo, IMPAQ International, LLC

With funding from the United States Department of Labor International Labor Affairs Bureau (USDOL/ILAB) Office of Child Labor, Forced Labor and Human Trafficking (OCFT) IMPAQ is conducting randomized controlled trials (RCTs) to evaluate the effectiveness of five child labor interventions in five countries - India, Malawi, Rwanda, Ecuador, and Costa Rica. The context of child labor varies in the five countries, from young children involved in mica mining in India, to tea plantation activities in Rwanda, working in tobacco farms in Malawi, and children leaving school and working primarily in the informal sector in a variety of urban environments in Ecuador and Costa Rica. The programs vary from educating and empowering the community to, facilitating returning to school, providing vocational training or intensive non-cognitive school curriculum modules to enable children to earn their secondary school certificate. Surveys are being implemented with 1) heads of household and women with children in India; 2) children aged 16-17 years in Rwanda in a vocational training program; 3) head of households in Malawi's tobacco growing area; 4) 15-21 years olds school dropouts in Ecuador enrolled in an intensive education program; and 5) children of all ages who have dropped out of school and are being enrolled back in Costa Rica. The questionnaires are designed to measure the impact of these programs. Questions include assessment of the prevalence of child labor. While the key research questions are shared across the five programs, the questionnaires are adapted to be relevant within the local context, the programs and populations. After the instruments are drafted, we will conduct cognitive testing with to refine the instruments. We will analyze the data from the testing paying special attention to change due to languages and cultures. We will analyze the differences in the ways respondents interpret and define concepts, and patterns in response and interpretation of response categories that might stem from cultural or linguistic factors. Once the instruments have been finalized based on the findings from the cognitive testing, we will pilot test each instrument with a sample that is similar to our target population. The pilot test will include some de-briefing questions for both the respondent and the interviewer. Analysis of these data will be conducted to assess if further changes are required in the instruments. We propose to present our analysis of these data from the cognitive and the pre-tests, and the resultant changes to the questionnaires.

A comparative youth survey in Azerbaijan, Georgia and Tajikistan. Results and experiences from multi-cultural questionnaire development and testing

Eliza Mandieva, University of Bamberg, Germany

We are implementing the international and interdisciplinary research project TEW-CCA in cooperation with universities and survey research institutes in Azerbaijan, Georgia and Tajikistan. The project aims at implementing empirical analyses of youth's transitions from education to work in the varying structural, institutional and cultural contexts of the Caucasus and Central Asia (CCA) region. The main innovative feature of the project is to conduct three large scaled, nationally representative standardized quantitative surveys in Azerbaijan, Georgia and Tajikistan. In each country it is planned to perform 2,000 standardized interviews. The survey mode is face to face interviews, PAPI. Adopting a life course transition approach, the sample unit comprises young people aged 18-35 years who have left the education system during the last ten years. The source questionnaire has been developed in accordance with questionnaire development guidelines on and experiences of school-to-work transition surveys for Western countries, but considering the specific national contexts of CCA countries. In this respect we relied on the expertise of national researchers from disciplines such as sociology, social-politics, economics and psychology. As a result the TEW-CCA team designed a questionnaire that provides answers reflecting specific research hypotheses on youth life course transitions in the CCA region. A source questionnaire is developed in English and target questionnaires will be produced in Azeri, Georgian, Tajik and Russian languages. Furthermore the target questionnaire will be in two languages for each country, national language + Russian. Comparability and quality of the data across countries despite their differences is one of the main goals to achieve. Accordingly a first part of the presentation is about the results - TEW-CCA Questionnaire , of the questionnaire development for surveys in the CCA Region; here emphasis will lie on kinds of questions and mix of ASQ and ADQ questions in the source questionnaire as well as on differences between the target questionnaires. First pre-tests will be conducted in spring 2016 in all chosen countries. For that reason the second part of the presentation will be dedicated to firsthand lessons learned/ experiences gained during the pre-tests.

Thursday, July 28th, 2:00 p.m. - 3:30 p.m.
Session: Resources for 3MC Surveys

Chair: Beth-Ellen Pennell, University of Michigan
Location: Great Lakes E

Recent developments in the standards for market, opinion, and social research of the International Organization for Standardization

Tom W. Smith, NORC at the University of Chicago

In 2002, the International Organization for Standardization (ISO) formed Technical Committee 225 to formulate standards for organizations and professionals conducting market, opinion and social research. TC225 came out with Market, opinion and social research -- Vocabulary and service requirements (ISO 20252) in 2006 and updated those in 2012. In 2009, TC225 created a new standard, Access panels in market, opinion and social research -- Vocabulary and service requirements (ISO 26362). Currently, TC225 is reviewing both standards with the object of folding the access panel standards into the general standards (ISO 20252). The TC is also about to issue a new standard for Digital analytics and web analyses in market, opinion and social research -- Vocabulary and service requirements (ISO/DIS 19731). It is likely that it will be finalized and issued in early 2017.

Cross-cultural survey guidelines: New and expanded

Beth-Ellen Pennell, University of Michigan
Kirsten Alcser, University of Michigan
Judi Clemens, University of Michigan
Julie de Jong, University of Michigan
Kristen Cibelli Hibben, University of Michigan
Mengyao Hu, University of Michigan
Jennifer Kelley, University of Michigan
Yu-chieh (Jay) Lin, University of Michigan

The number and scope of studies covering many cultures, languages, nations, or regions have increased significantly over the past decade. This has led to a growing need to provide information on best practices across the multiple aspects of cross-cultural survey design and implementation to ensure the collection of high-quality comparative data. The Cross-Cultural Survey Design and Implementation (CSDI) guidelines initiative was created to address this gap. Specifically, the aim of the initiative was to develop and promote internationally recognized guidelines that highlight best practice for the conduct of comparative survey research across cultures and countries. The intended audience for the guidelines includes researchers and survey practitioners planning or engaged in cross-cultural or cross-national research. The guidelines were first published in 2008, with several updates in the interim. Over the past two years, 35 faculty, staff and students from 13 organizations across the globe have worked to update and expand these guidelines. In addition to updating the literature, content has expanded to cover mixed mode and design considerations, emerging technologies, paradata, and data analysis approaches and tools.

The guidelines cover all aspects of the survey life-cycle and include the following chapters: Study Design Considerations; Study and Organizational Structure; Tenders, Bids and Contracts; Ethical Considerations; Sample Design; Questionnaire Design; Translation; Adaptation; Survey Instrument Technical Design; Pretesting; Interviewer Recruitment and Training; Data Collection Implementation, Collection and Analysis of Paradata; Harmonization of Survey and Statistical Data; Data Processing and Statistical Adjustment; Dissemination of Survey and Statistical Data; Assessing Quality for Cross-Cultural Surveys; and Data Analysis: Approaches and Tools. The guidelines can be found at http://ccsg.isr.umich.edu.

International program in survey and data science
Frauke Krueter, University of Maryland

The demand for survey statistics and survey methodology continues to be strong, despite the increasing interest in Big Data. In fact, recent trends seem to indicate that large companies that deal with data but are not in the business of surveys are beginning to build their own in-house survey operations. For this reason we designed a new international modularized continuing education program in which learners across the world can participate. The program combines asynchronous and synchronous training elements designed by leading experts for each particular topic. Curricular corner stones are courses related to research design, data collection, data curation and storage, data analysis, as well as data output and privacy. At each step, critical features with respect to survey data are considered, as well as the integration of survey data, administrative data, and other found online data. The presentation will describe the structure of the program, highlight the international partners that already signaled strong interest in participating, and describe the curriculum in more detail. We hope to engage with attendees in a discussion of training needs in their particular work environments.

Thursday, July 28th, 2:00 p.m. - 3:30 p.m.
Session: TSE Paradigm for Cross-cultural Surveys
Chair: Zeina N. Mneimneh, University Michigan
Location: Great Lakes A/B

*** Improving cross-national/cultural comparability using the total survey error paradigm**
Tom W. Smith, NORC at the University of Chicago

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.

*** Organizing and managing comparative surveys in a total survey error perspective**
Peter Ph Mohler, COMPASS & Mannheim University

Talking on organizing and managing comparative (3MC) surveys beyond episodical insights is a challenge. That is because of a lack of systematic literature on management and organizational issues compared to the wealth of literature on other methodological issues such as design, sampling, translation, or even contracting.
However, the impact of different management or organizational forms on survey quality is not self-evident. Firstly, one can ask: what is their on a 3MC survey and which stages of the survey production process can be affected? Secondly, if there are impacts, is there a scientific approach that allows identifying adverse impacts and, best, helps to remedy such impacts. As simple as these questions are as obvious it is that simple solutions seldom are in general as well in comparative survey management. A paper given at this point in time will and can not offer cookbook like recipes. The diversity of today's survey techniques and practices is but one major obstacle. Consider a, say thirty

country simultaneous survey, where respondents reply in self completion and interviewer driven, paper and pencil, computer assisted, hand held devices and what else mode comes to ones mind are used. Each of them might require country specific optimal organization and management. And there are many more issues like that to observe in 3MC surveys. Despite such diversities, there are overarching management principles that can help to guide organization and management. Two will be dealt with in the paper:

1. Insights from cognitive research on management
2. Introducing a Total Survey Error perspective

Cognitive research on decision making had a major impact on how companies are organized today. Key concepts are self deception (i.e. beliefs in ineffective management habits) or concentration on easy solutions (i.e. avoiding difficult management decisions). The Total Survey Error perspective links survey research with modern high quality production processes and introduces quality management tools into the survey production process.

Migration and pensions: How to collect valid data on pension entitlements of immigrants

Thorsten Heien, TNS Infratest Sozialforschung
Dina Frommert, Deutsche Rentenversicherung Bund

Collecting valid data on pension entitlements of immigrants is increasingly important for the monitoring and further development of pension systems, regardless of recent migration dynamics. On the one hand, immigrants are disadvantaged with respect to sociodemographic and job characteristics (e.g. education, skill-level, employment patterns, income) resulting in lower old-age incomes than the native population. On the other hand, current data and its quality are inadequate: There is a general lack of appropriate indicators in official statistics while in survey research immigrants are often underrepresented due to limited language skills. The latter point results in smaller sample sizes of immigrants being too small for statistical testing and , even more problematic , their survey representation is systematically distorted. The issue is complicated further by the need to collect data on possible pension entitlements in other countries. This is not only relevant for immigrants but affects the native population as well, as globalization processes result in increasingly transnational work histories. The paper discusses the question of how to collect valid data on pension entitlements using the example of the newly implemented survey on Life-courses and old-age provision (Lebensverlaufe und Altersvorsorge; LEA). LEA will collect more than 10,000 interviews (CAPI) of people living in Germany who are 40 to 60 years old in 2016 and link these individually to administrative data from respondents' statutory pension accounts. The paper is guided by the total survey error approach (TSE), an attempt to conceptualize and generalize errors of sample survey statistics. According to TSE, two separate inferential steps are required in any survey: the first one is from the (edited) response to the underlying construct of interest (measurement) where relevant concerns are processing error, measurement error and validity. The second step is from an estimate based on a set of respondents to the target population (representation) where the relevant error sources are nonresponse error, sampling error and coverage error. The paper discusses the different measures introduced in LEA to reduce these potential errors with a focus on immigrants and pension entitlements (e.g. sample design, questionnaire design, translation issues, interviewer training, respondent incentives), and presents first fieldwork results.

TSE and survey quality in 3MC Surveys

Beth-Ellen Pennell, University of Michigan
Kristen Cibelli Hibben, University of Michigan
Lars Lyberg, Stockholm Univeristy
Peter Ph. Mohler, COMPASS & Mannheim University
Gelaye Worku, Stockholm University

Multinational, multiregional, and multicultural (3MC) surveys are becoming increasingly important to global and regional decision-making and theory-building. With this has come renewed awareness of the importance of survey data quality and comparability. The TSE framework helps to organize and identify error sources and to estimate their

relative magnitude. This TSE approach can assist those planning comparative surveys to evaluate design and implementation tradeoffs. In this presentation, we introduce a TSE framework adapted and expanded from Groves et al. (2009), Tourangeau et al. (2000) and Smith (2011) for comparative 3MC survey research. The proposed framework integrates error sources with methodological and operational challenges that are unique to or may be more prominent in 3MC surveys. We also integrate the dimensions of cost, burden, professionalism, ethical requirements, and other design constraints.

In addition to TSE, we discuss how "fitness for intended use" and survey process quality offer two additional and important approaches through which to understand and assess 3MC survey quality. Fitness for intended use provides a general framework for assessing the quality and defines the essential dimensions of quality including comparability, coherence, relevance, and accuracy (i.e., TSE). A third important aspect of quality is survey process quality management, and the notion of continuous process improvement. 3MC survey organizations vary in the processes they generally monitor for quality purposes. We present examples of how, if minimum standards and quality guidelines can be established, monitored and assessed, the quality of each survey can be measured based on quality indicators from each organization, thereby creating a quality profile that fully documents outcomes and allows end users to assess overall survey data and comparability.

Thursday, July 28th, 4:00 p.m. - 5:30 p.m.
Session: Developing and Testing Iterative Race Question Stems and Instructions for Multicultural, Hard-to-Count American Indian and Alaska Native Populations for the 2020 Census

Chair: Laurie Schwede, U.S. Census Bureau
Location: Great Lakes D

Qualitative research on questions of race, tribe identification, and tribal enrollment for multicultural American Indians

Rodney L. Terry, U.S. Census Bureau
Leticia Fernandez, U.S. Census Bureau
Laurie Schwede, U.S. Census Bureau
Aleia Fobia, U.S. Census Bureau

Race and ethnicity data are typically collected in demographic surveys and censuses in the U.S., and are considered by various stakeholders to be important for describing the U.S. population and enforcing U.S. anti-discrimination laws. However, the American Indian or Alaska Native (AIAN) race category poses a unique situation to respondents because this category, as defined by U.S. Office of Management and Budget (OMB), includes the provision of maintaining tribal affiliation or community attachment. The 2010 U.S. Census race question had a checkbox for the AIAN category, as well as an instruction to print the name of the enrolled or principal tribe in a write-in space. Previous qualitative research and consultations with AIAN stakeholders have shown that many American Indians find the write-in instruction confusing to respondents who have complex relationships to multiple tribes, or who think this instruction is only asking for enrolled tribe relationships. As part of addressing these challenges, we present findings from three 2020 U.S. Census research projects with American Indians as part of a contributed session entitled Developing and Testing Iterative Race Question Stems and Instructions for Multicultural, Hard-to-Count American Indian and Alaska Native Populations for the 2020 Census. To make AIAN race and tribe reporting less confusing, the U.S. Census Bureau conducted two focus group studies and one cognitive interview study to develop improved AIAN write-in line instructions and a new and separate tribal enrollment question. We identify findings that arose for the American Indian participants, including the significance of enrollment status when reporting tribe information, and which AIAN category or write-in line instruction best allows them to self-identify their race. Also identified are opinions about the Census Bureau collecting data on tribal enrollment and preferred tribal enrollment question

wordings. Finally, the implications of this research for multicultural and multiregional research are discussed. Results of the focus group and cognitive testing research informed which instructions and AIAN categories wordings were chosen for a national, split-ballot test in 2015. The research on a possible tribal enrollment question will help inform the development of a new tribal enrollment question for inclusion in another national, split ballot test in 2017. The results of these national tests will help inform which content are chosen for inclusion in the 2020 U.S. Census.

Overview of Three qualitative Census research projects on race and tribal enrollment questions with American Indian/Alaska Native subpopulations: American Indians, Alaska natives, and Latin American indigenous groups

Laurie Schwede, U.S. Census Bureau
Rodney Terry, U.S. Census Bureau
Aleia Clark Fobia, U.S. Census Bureau
Leticia Fernandez, U.S. Census Bureau

Developing standardized census questions that different segments of multicultural populations understand and answer as intended poses challenges for survey researchers. The American Indian or Alaska Native (AIAN) race category is one such multicultural population encompassing American Indians, Alaska Natives, and Central and South American indigenous groups who maintain tribal affiliation or community attachment (Office of Management and Budget, 1997). Three recent qualitative research projects in preparation for the 2020 U.S. Census have focused on issues related to race for American Indians and Alaska Natives in these three subpopulations. The first consisted of focus groups to learn how participants answer the race question and determine which of six alternative race question/instruction formats should go into a 2015 national split-panel content test. The second project involved cognitive testing of three new race question formats with revised examples and special subpopulation checkboxes, also for the 2015 test. The third is using focus groups and cognitive interviews to develop and test a new tribal enrollment question for a 2017 split-panel test and possible inclusion in the 2020 Census. This is the proposed final presentation in the accepted 3MC contributed session, Developing and Testing Iterative Race Question Stems and Instructions for Multicultural, Hard-to-Count American Indian and Alaska Native Populations for the 2020 Census. The first three papers each present results across the three studies for one subpopulation: American Indians (Terry); Alaska Natives (Clark Fobia); and Latin American indigenous groups (Fernandez). This synthesizing overview paper summarizes similarities and differences across these three distinct subpopulations and three projects in response patterns, degree of cultural fit, and participant preferences. It identifies factors affecting the extent to which the race, tribe and enrollment questions were understood and answered appropriately: 1) complexity of racial or tribal affiliation (all groups); 2) tribal enrollment status (American Indian); 3) lack of familiarity with or non-use of the concepts of tribe and enrollment (Alaska Natives and indigenous immigrants), and 4) translation and formatting issues (Latin American indigenous immigrants). The tribe and enrollment concepts are based on American Indian concepts and practices and appear to be more easily understood by them, while the concepts appear to be more ambiguous to Alaska Natives and unfamiliar to some Latin American indigenous immigrants. Recommendations are offered for improving the race and enrollment questions for AIAN and lessons learned are identified. The paper concludes with wider survey methodology implications for developing and testing questions for other multicultural and hard-to-count populations.

Testing potential 2020 Census race question wording and instructions with Latin American indigenous immigrants

Leticia Fernandez, U.S. Census Bureau
Aleia Clark Fobia, U.S. Census Bureau
Rodney Terry, U.S. Census Bureau
Laurie Schwede, U.S. Census Bureau

In order to monitor and address disparities in a broad range of outcomes for populations that historically have experienced discrimination, the Census Bureau and other agencies are required by the Office of Management and Budget (OMB) to collect data on self-reported race and ethnicity. The racial categories defined by OMB are: American Indian or Alaska Native (AIAN), Asian or Pacific Islander, Black, and White. The AIAN race category encompasses indigenous peoples from North, Central and South America who maintain tribal affiliation or community attachment. In 2010, 23 percent of those who reported as AIAN also identified as Hispanic. Hispanic American Indians include U.S.-born people as well as immigrants who identify as members of indigenous groups in Latin America. Latin American indigenous groups living in the United States are hard-to-count populations particularly vulnerable to discrimination. They are racial and ethnic minorities characterized by a high proportion of undocumented immigrants. Given the history of marginalization and discrimination of indigenous groups in Latin America, they are also likely to have very low levels of education and experience high rates of poverty. In addition, they may not be proficient in English or, in some cases, Spanish, and may not identify with the overarching indigenous category; rather, they may report their language or place of origin as their main identifier. This presentation focuses on Spanish-speaking Latin American indigenous groups as paper 3 in the contributed session, Developing and Testing Iterative Race Question Stems and Instructions for Multicultural, Hard-to-Count American Indian and Alaska Native Populations for the 2020 Census. We report findings from two multicultural studies conducted in English and Spanish under the 2020 U.S. Census Research Program. These studies were designed to test various race question wording options to improve reporting under the AIAN category. In the first of these studies, focus groups were conducted to learn how Latin American indigenous immigrants and other American Indian populations answer the race question, and to select a format for further testing in a nationally representative field test in 2015. The second study involved cognitive interviews to test new AIAN race question formats and examples. We identify and discuss issues with translation, formatting, examples, and lack of familiarity with constructs and concepts of tribal enrollment in the U.S. as potential barriers to identification as AIAN. We conclude by discussing the implications of this research for multicultural and multiregional studies.

Examples, instructions, and belonging: Alaska native social organization and the race question

Aleia Clark Fobia, U.S. Census Bureau
Laurel Schwede, U.S. Census Bureau
Leticia Fernandez, U.S. Census Bureau
Rodney Terry, U.S. Census Bureau

In the 2010 Census, 5.2 million people in the United States identified as American Indian and Alaska Native (AIAN) either alone or in combination with one or more other races. Alaska Natives, people indigenous to Alaska, are a subset of those who identified as AIAN in 2010. In the U.S., one of the reasons race data is collected is to ensure that the government can track disparities that continue to exist along racial lines. The 2010 Race question included an instruction to print the name of the enrolled or principal tribe. Previous research has shown that while American Indians are generally familiar with writing the names of their tribes, Alaska Natives have difficulty with the concepts of enrolled or principal tribe. This presentation focuses on Alaska Natives as paper 2 in the contributed session, Developing and Testing Iterative Race Question Stems and Instructions for Multicultural, Hard-to-Count American Indian and Alaska Native Populations for the 2020 Census. Similar to American Indians, the U.S. federal government recognizes 229 Alaska Native tribal entities. The entities that have become federally recognized tribes include a vast array of organizations that include tribal governments and councils, native villages, and community associations. In addition to these entities, because of the 1971 Alaska Native Claims Settlement Act (ANCSA), Alaska Natives are also enrolled as shareholders in 13 Alaska Native Regional Corporations that administer land and financial claims. This multi-layered system of association and belonging complicates issues of identity for Alaska Natives and results in confusion when asking them to report their enrolled or principal tribe. In this presentation, we report findings specific to Alaska Natives (AN) from three multi-cultural research studies. In the first study, two AN focus groups were conducted to test alternative race question wordings. In the second study, cognitive interviews were conducted with a diverse group of Alaska Native participants to further test alternative race questions and their formatting. In the third study, we are conducting additional focus groups with Alaska Native participants, followed by a round of

cognitive interviews to inform a new question on tribal enrollment. We identify issues with certain examples and instructions that are particularly problematic for Alaska Native's racial self-identification. We discuss how multi-layered systems of belonging in Alaska Native social organization add complexity to wording of a tribal enrollment question. We also identify implications and takeaways for wider issues in question design in multicultural and multinational contexts.

Thursday, July 28th, 4:00 p.m. - 5:30 p.m.
Session: Innovative Tools/Methods for Data Collection

Chair: Yu-chieh (Jay) Lin, University of Michigan
Location: Great Lakes A/B

Facebook advertisements for cross-cultural survey recruitment: Insights from the 42-country World Relationships Study

Robert Thomson, Hokkaido University, Japan
Masaki Yuki, Hokkaido University, Japan

With over 80% of Facebook's 1.3 billion active users hailing from countries outside of the United States, survey recruitment via Facebook advertisements suggest unprecedented potential for the cross-cultural researcher conducting multi-country comparative studies. However, how can one capture the elusive attention of Facebook users and persuade them to engage in academic surveys? Drawing on the concept of functional sources of attitudes (Katz, 1960), we developed an online survey to be administered in 21 languages across 40 countries, which sought to tap into Facebook users' intrinsic motivations of making sense of their environment and relationships (knowledge function) and the motivation of bolstering one's sense of self-worth (value-expressive function) (Daugherty et al., 2005): in addition to containing a key 12-item scale central to the study, the survey 1) included scales associated with close relationships, and 2) offered immediate feedback and 'scores' based on participants' responses, with comparisons to previous participants' aggregate scores. In order to recruit participants to the survey, Facebook ads were deployed, targeting Facebook users 18 years old or over, in 42 countries. No extrinsic motivators such as raffles or cash incentives were offered. The online survey, with an average response time of 13 minutes, garnered almost 15,000 valid responses, with at least 300 valid responses per country, however participation was heavily biased towards female respondents. Response rates, cost-performance, and data quality issues will be discussed, and comparisons will be made with the authors' previous attempts at using Facebook advertising cross-culturally, in which extrinsic incentives (gift voucher raffle) were the central motivators. This presentation offers general online researchers insight - both the opportunities and pitfalls - associated with the effectiveness of recruiting participants to online surveys using Facebook advertisements, across a globally diverse set of countries.

Developing a mobile fieldwork management and monitoring system: Challenges of cross-national implementation

Elena Sommer, European Social Survey, City University London
Sarah Butt, European Social Survey, City Univeristy London
Lennard Kuijten, CentERdata, Tilburg University

Cross-national surveys without centrally managed fieldwork (such as the European Social Survey (ESS)) are faced with survey agencies in different countries using different methods and systems to manage and monitor their fieldwork. Many countries do not have access to their own electronic fieldwork management system and are still reliant on the completion of paper records. This can result in delays and inconsistencies in the flow of information between the fieldwork agencies and central survey team making it difficult to monitor fieldwork in an effective, consistent and timely manner across countries. As part of the EU-funded DASISH and SERISS projects, researchers from the ESS, SHARE and CentERdata have been working on developing an electronic fieldwork management and monitoring system (FMMS) for use in all ESS countries. The tool consists of a centralised case management system and a mobile

app which interviewers will use on the doorstep via a handheld device such as a tablet or smartphone to record information about contact attempts and interview outcomes, replacing the existing paper contact forms. The tool is intended to ensure that interviewers, survey agencies and the central survey team have access to accurate, up to date information throughout fieldwork enabling them to manage fieldwork progress more effectively. Despite the clear potential benefits of such a fieldwork management and monitoring tool, there are, however, several challenges for its implementation cross-nationally. In early 2016 a scoping exercise was conducted among key ESS stakeholders, including the survey agencies carrying out ESS fieldwork, to explore potential difficulties such as transfer of information across countries, legal restrictions concerning data protection, security, ownership and storage in the central server as well as resources and costs associated with availability of mobile technology and IT support. There are also issues associated with developing a tool that is sufficiently flexible to ensure compatibility with agencies' existing sample management systems and variations in sample information across countries. In our presentation we will first briefly demonstrate the main features of the FMMS tool including the app prototype. We will then go on to present the results from the scoping study, discussing the major challenges to implementation identified and issues raised by stakeholders in different European countries. We will end by considering the implications of these findings for a future roll out of the proposed fieldwork management and monitoring tool on a decentralised cross-national survey such as the ESS.

Innovative data collection methods

Utz Pape, World Bank
Johan Mistiaen, World Bank

Data collection in fragile countries faces two major constraints. First, field access is often limited to monitor data collection. This puts data quality at risk if enumerators cannot be sufficiently supervised. Second, the context is highly volatile demanding timely data collection with swift results. The note presents three measures to overcome these barriers and deliver high quality household consumption data using tablet face-to-face interviews. The three innovations go beyond traditional tablet survey collection by an innovative online management of the data collection, dynamic on-the-fly data validation and real-time monitoring and analysis of the data.

Framework to assess the maximum expected response rate for different survey designs and field conditions

Francois Laflamme, Statistics Canada
Sylvie Bonhomme, Statistics Canada

Statistics Canada, like many national statistical organizations, has observed a downward trend in response rates, particularly since 2011. Changes in the external environment (e.g., increased number of cellular phone only households, increased use of telephone caller display and new technologies) as well as internal structural changes (e.g., the introduction of Electronic Questionnaire (EQ) collection mode and introduction of the new Common Household Frame) have led to a sustained decrease in response rates for household surveys. To address this complex issue, the agency is currently looking at different initiatives and options to improve data collection processes, and is formulating strategies to improve response rates. In the meantime, recent surveys have also provided a better understanding of external and internal factors that might impact response rates before and during data collection. Data collection organizations have no or little control over some of these factors, while for other factors they have some control, especially during collection. Thus, before collection starts, there is a maximum/optimum attainable response rate that can be achieved by the data collection organisations given all the factors for which data collection organizations have no or little control. The main objective for each data collection organization in that context is to draw near to this maximum response rate using the best data collection practices before and during collection. This paper begins with an overview of the factors impacting response rates before and during collection. The next two sections describe in more details the factors impacting response rate before and during collection. In both sections, these factors are divided into two categories: factors for which data collection organisations have no or little control

and factors for which they have some control. The next part presents the main challenges facing data collection organisations during collection to obtain response rates that are as close as possible to the maximum achievable response rate. Finally, the last section discusses the relationship that exists between the maximum and observed response rates in the perspective of developing an indicator to assess the overall performance of the data collection organization.

Thursday, July 28th, 4:00 p.m. - 5:30 p.m.
Session: Interviewer/Respondent Interactions

Chair: Yfke Ongena and Marieke Haan, University of Groningen
Location: Michigan Ballroom II

Examining respondent-interviewer interactions using behavior coding data and paradata

Allyson L. Holbrook, University of Illinois at Chicago
Timothy P. Johnson, University of Illinois at Chicago
Young Ik Cho, University of Wisconsin, Milwaukee
Sharon Shavitt, University of Illinois
Noel Chavez, University of Illinois at Chicago
Saul Weiner, University of Illinois at Chicago

This presentation reports on analysis of behavior coding and paradata from in-person interviews in which members of four different racial and ethnic groups (non-Hispanic White, non-Hispanic African-American, Mexican-American, and Korean-American) were interviewed by an interviewer who was matched on race/ethnicity. Half of the Mexican-American respondents were interviewed in English and half in Spanish. Half of the Korean-American respondents were interviewed in English and half were interviewed in Korean. All interviews were audio and video-taped and paradata (including question reading and response latencies) was collected. Respondents' and interviewers' verbal behaviors were coded by coders listening to the audio recordings of the interviews. We use these measures to assess whether there are differences in interviewer-respondent interactions across racial and ethnic groups.

Answers to standardized questions: Recognizing codable answers and maintaining standardization and rapport

Nora Cate Schaeffer, University of Wisconsin Survey Center
Jennifer Dykema, University of Wisconsin Survey Center
Dana Garbarski, Loyola University Chicago

Standardized interviewing aims to make interviewers interchangeable , ideally all standardized interviewers will behave the same way in a given situation -- to increase the reliability of measurement. Key to the implementation of standardization is the concept of a codable answer. If the respondent provides a codable answer, the interviewer moves on to the next question. If the respondent does not, the interviewer must follow-up. Central though this concept is, the field lacks widely accepted conceptual and operational definitions. To develop such guidelines, we need an analysis of how respondents answer different forms of survey questions (that is, formatted for different types of responses), how respondents use conversational elements to supplement or embellish codable answers, and instructions for interviewers faced with these answers. The challenges of understanding the structure of answers are more complex in a cross-cultural setting, although some features of turn construction are shared across many cultures and languages (Stivers et al. 2009). We describe the relationship between question form and respondents' answers in several US studies to identify some of the components of answers that a broader cross-cultural analysis might explore. For example, respondents may provide reports of various types (Schaeffer & Maynard 1996, 2008) if their candidate answer does not fit the assumptions of the question or response categories. Respondents also use other common conversational practices that are not considered in interviewer training. For instance, instead of directly stating yes or no to a yes-no question form, respondents may repeat part of the question (e.g., I did) or

provide a synonym that has not been explicitly defined as codable or not (e.g., yep or probably). Respondents may also add modifiers (e.g., about) and other elaborations to a codable answer. These conversational practices pose challenges to interviewers who must balance the needs of standardization , which might require following up an answer of probably with Would you say 'yes' or 'no'? , and those of rapport , which recommend not following up because probably likely means probably yes in ordinary conversation (Garbarski, Schaeffer, & Dykema forthcoming). In this presentation we 1) describe the relationship between question form and the way respondents answer, 2) provide an operational definition of a codable answer, 3) catalogue varieties of conversational elements respondents add to or treat as synonyms for codable answers, and 4) offer guidelines for when interviewers should follow-up versus display their understanding of the respondent's answer by proceeding to the next question.

*** Using behavior coding to evaluate question comparability**

Timothy P. Johnson, University of Illinois at Chicago
Allyson Holbrook, University of Illinois at Chicago
Young Ik Cho, University of Illinois at Chicago
Sharon Shavitt, University of Illinois at Chicago
Noel Chavez, University of Illinois at Chicago
Saul Weiner, University of Illinois at Chicago

Initially developed to evaluate interviewer performance, behavior coding has also been used effectively to investigate respondent processing of survey questions. This chapter will provide an overview of the use of behavior coding to examine the issue of cross-cultural comparability of cognitive assessments of survey questions. We begin with a comprehensive review of previous research on this topic, followed by original analyses of survey data from three projects. These analyses will focus on comparisons of behavior coding measures of question comprehension and response mapping across multiple race/ethnic groups in the U.S., and on the question of the cross-cultural validity of behavior coding as a method for examining respondent cognitive processing of survey questions. These analyses will be based on cross-classified, multi-level models that examine and control for variables measured at the respondent, question, and response levels of analysis. Our discussion will focus on the unique aspects of using behavior codes to detect cultural variability in question processing, as well as limitations, and make recommendations for future research directions.

Thursday, July 28th, 4:00 p.m. - 5:30 p.m.

Session: Methods to Support the Development of Cross-cultural Source Questionnaires

Chair: Michèle Ernst Stähli, FORS
Location: Michigan Ballroom I

Challenges of measuring the attitudes towards the scope of government cross-nationally: Evidences from international surveys in Venezuela

Roberto Briceno-Rosas, GESIS - Leibniz Institute for the Social Sciences

Governments and their scopes have developed differently around the world. Therefore cultural particularities are to be expected. If respondents understand key concepts related to the scope of the government differently as intended by researchers, we can assume problems achieving comparable measurements of attitudes. Understanding the cultural particularities is therefore crucial for understanding cross-cultural data. In this paper, I examine the concepts used in survey questionnaires of the International Social Survey Programme (ISSP) for measuring the attitudes towards the scope of government. I focus on the Venezuelan context and make use of cognitive interviewing to reconstruct respondents' understanding of the concepts within a survey environment. The results are used to generate explanations for the particularities of Venezuelan survey data and to highlight the challenges of comparable measurement in this research field.

More than language: Strategies for detecting, resolving and coping with issues of cross-cultural transferability of concepts

Michael Ochsner, FORS
Michele Ernst Staehli, FORS
Karin Nisple, FORS
Alexandre Pollien, FORS
Marlene Sapin, FORS

International comparative surveys aim at achieving comparable results across different countries. In order to be able to compare results across national contexts, the concepts under scrutiny must be meaningful and valid in all surveyed nations. This, however, is a real challenge. In our presentation, we discuss issues of transferability of concepts between cultures and present strategies to detect and resolve such problems. Using data from Switzerland, it is possible to analyse the cross-cultural transfer of concepts between countries (when applying a foreign concept) as well as within a country (when transferring a concept from one language region to another). We use two concepts that are frequently used in the scientific literature in their cultural contexts of origin but proved to be difficult to implement in the other cultural context we tried to transfer the concept to. Ethnic group membership serves as an example for the cross-cultural transfer between countries. We use data from a pilot study for the ISSP 2015 administered in 2013 in the German part of Switzerland. While this concept is well known in the Anglo-Saxon countries, it is not so much used in continental European countries like Switzerland. The concept secondo serves as an example for a cross-cultural transfer within a country. We use data from cognitive tests administered in the French part of Switzerland in 2014 while preparing a survey module for Switzerland. The concept secondo is widely known in Germany and the German part of Switzerland referring to second-generation immigrants. However, the concept is not known among people living in the French part of Switzerland. The results reveal that the concepts cannot be easily transferred from one cultural context to another even though they are scientifically useful in their cultural context of origin. Thus, survey designers should be aware of transferability issues while constructing questionnaires. In our presentation, we will tackle such problems and stress that issues of comparability cannot always be resolved by optimizing translation but need much more scrutiny and might even lead to a revision of the concepts as such. We provide strategies and methods to detect, resolve and cope with transferability issues.

Can advance translation contribute to identifying source questionnaire problems at the concept level?

Brita Dorer, GESIS-Leibninz-Institute for the Social Sciences

In cross-cultural surveys, typically an English source questionnaire is translated into various target language versions. It is highly important that the resulting translations are as comparable as possible in order to receive comparable data. The European Social Survey (ESS) is a biennial social sciences survey fielded in 25+ countries since 2002. In order to achieve the best-possible quality of its questionnaire translations, it has used the TRAPD process since its first round. As also the source text matters for the translation quality, in 2009, it started carrying out systematic 'advance translations' in order to be able to detect source questionnaire problems before finalising the source text. Translating a pre-final version of the source questionnaire into different languages, applying the usual 'team approach', is used as method to detect problems in this version of the source questionnaire. The findings are fed back to the source questionnaire developers and considered when finalising the source text. While the issues detected by advance translation are typically on the linguistic level, in many cases also problems at the intercultural and/or content level are identified. The paper will briefly describe the method of advance translation as carried out in the ESS. It will then present and discuss examples where advance translation has led to detecting source questionnaire problems at the concept level. These problems will be discussed together with the solutions found in the final survey instruments. The goal will be to find out to what extent advance translation is a useful tool to detect and tackle problems in a cross-cultural survey at the concept level.

Thursday, July 28th, 4:00 p.m. - 5:30 p.m.
Session: New Sage Handbook of Survey Methodology
Chair: Christof Wolf, GESIS
Location: Great Lakes E

Introducing the Sage Handbook of Survey Methodology: State and future of survey methodology
Christof Wolf, GESIS
Dominique Joye, University of Lausane
Tom W. Smith, NORC at the University of Chicago
Yang-chih Fu, Academia Sinica, Taipei

Survey methodology, i.e. the scientific approach to surveys, is an advancing scientific field. Once part of social sciences' training in research methods survey methodology is emerging as a discipline in its own right. The central aim of the Sage Handbook of Survey Methodology is to consolidate scattered knowledge and to provide a comprehensive overview of this field. Its 43 chapters are organized around three principles: First, the flow of chapters roughly follows the stages in the survey lifecycle. Second, the Total Survey Error framework serves as common point of reference for the contributions. Third, the challenges posed by comparative, cross-cultural survey research are extensively discussed for each stage of the survey lifecycle. 73 experts from 16 countries in Europe, North America and Asia have contributed to the handbook which represents the state of the art of survey methodology and in particular comparative survey research.
This session will be organized around three questions, each introduced by a short intervention by an editor of this handbook, which we want to discuss with the audience:

- To what extent is survey methodology a discipline? What is its position in the landscape of the social sciences in particular and in the broader field of scientific disciplines?
- Why should survey methodology be comparative? Could a comparative perspective be even useful for designing, implementing and analysing a survey in a one-country study?
- Finally, how could the ever growing number of data sources be used in conjunction and combined with surveys? In other words: how does "big data" challenge survey research, in particular when taking into account a comparative perspective?

Friday, July 29th, 9:00 a.m. - 10:30 a.m.
Plenary: Evaluating and Improving Data Integrity
Chair: Patty Maher, University of Michigan
Location: Great Lakes Ballroom

Preventing interview fabrication
Patty Maher, University of Michigan
Jennifer Kelley, University of Michigan
Beth-Ellen Pennell, University of Michigan
Gina-Qian Cheung, University of Michigan

This panel presentation will explore innovative approaches to preventing interview fabrication. The diffusion of new technologies in survey research has led to new and innovative approaches to quality control. These procedures are being implemented to not only address overall data quality but to also make it more difficult to fabricate all or part of an interview. In this presentation, we draw upon examples from large scale complex surveys in a variety of challenging international contexts. The methods described go well beyond traditional call-back verifications to a sample of interviewer's cases. These new approaches include interviewer supervision models, analysis of tailored reports that combine rich paradata with survey data in 'real time', comparisons with previous waves of data

collection (where present), extensive use of paradata and audio-recording interviews to prioritize evaluation and verification of cases, use of biometrics such as fingerprints, digital photography, and use of GPS, including live tracking of interviewer travel in and among sampled segments.

* New frontiers in preventing data falsification in international surveys

Michael Robbins, Arab Barometer
Noble Kuriakose, Survey Monkey

Concerns about data falsification in survey research are as old as the field itself. Cheating will continue to exist as long as it seems less costly than faithfully carrying out the survey to interviewers, supervisors, or the survey firm. In recent years, tests designed to evaluate whether data fabrication has occurred suggest that cheating remains a significant problem, especially in international contexts where data collection often occurs face-to-face. We provide an overview of new approaches for detecting potential fraud, arguing that these checks should become standard practice for identifying and investigating cases of likely falsification. We also make the case that further steps, including investments in newer data collection technologies and closer on-the-ground monitoring, are needed to identify fraud in real time when steps can be taken to correct the survey. Finally, we argue that organizations engaged in international survey research must be much more transparent and collaborative. Transparency and collaboration will incentivize local firms to improve the quality of their research practices and take steps to dis-incentivize cheating at the local level.

Evaluating data quality in international surveys: A multi-dimensional approach

Katie Simmons, Pew Research Center
Steve Schwarzer, Pew Research Center
Gijs van Houten, Pew Research Center
Courtney Kennedy, Pew Research Center

Ensuring data quality in face-to-face surveys is a challenge given limited ability to monitor interviewer and supervisor activity in the field. This difficulty is compounded for international surveys when those who commission the research may not even be in the same country as the local vendor and are thus one more step removed from fieldwork. Computer-assisted personal interviewing (CAPI) presents a promising improvement for oversight in face-to-face surveys, but also comes with its own difficulties. Despite a vendor's best efforts to ensure accurate data collection through these devices, it is still possible to encounter unreliable time measurement, inconsistent connectivity, and inaccurate measurement of geolocations, among other problems. Given these challenges, we explore a multi-dimensional approach to evaluating data quality in international face-to-face surveys that relies on analysis of both substantive data (such as duplicate responses, item non-response patterns, straightlining, etc.) and available paradata (such as time of interview, time between interviews, interviewer workload, etc.). Using a set of nationally representative international surveys, we employ this ex-post approach to identify potential problems with the data, ranging from basic human error to poor interviewing practices to suspected falsification. We validate the approach by evaluating other data quality measures in the survey as well as patterns of response on substantive questions. The paper discusses the sensitivity of using a single measure to identify suspicious cases, the robustness of the multi-dimensional approach, and the limitations of relying on imperfect paradata. The paper proposes some possibilities for future research on evaluating data quality in international surveys.

Interviewers' deviations in face-to-face surveys: Investigations from experimental studies

Natalja Menold, GESIS

In face-to-face surveys the interviewer is a key actor, who may affect the quality of survey data. There have been some studies, which address interviewers' impact, for example, by evaluating interviewer variance in secondary analyses. However, there is a lack of knowledge on how falsifications may impact interview data as well as which working conditions influence the results of interviewers' work. Within the research conducted in collaboration with

Prof. Peter Winker (University of Giessen), which was founded by German Research Foundation, identification of interviewers' falsifications and possible effects of interviewers' work organization on the accuracy of the interviewers' results were investigated. In the first experimental study indicators based on specific properties of falsified interview were developed, tested and used as a relevant part of the multivariate method for identification of falsified data. The results show that particularly differences in response sets and specific patterns of response behavior could be effectively used to identify falsifications. In the second experimental study the impact of payment scheme, instructions and task difficulty on the accuracy of interviewers' work was analyzed. The results show that there were lower deviations if the interviewers were paid per time, while the variation in the instruction did not have an impact. In addition, there were more deviations by interviewers in the case of break offs, which was associated with a high task difficulty. The results are discussed with respect to the prevention and detection of interviewers' deviations.

Friday, July 29th, 11:00 a.m. - 12:30 p.m.
Plenary: Prevailing Issues and the Future of Comparative Surveys

Chair: Timothy P. Johnson, University of Illinois at Chicago
Location: Great Lakes Ballroom

*** Prevailing issues and the future of comparative surveys**

Lars Lyberg, Stockholm University
Lilli Japec, Statistics Sweden
Can Tongur, Statistics Sweden

The interest in comparisons between countries, regions, and cultures (3MC) has increased during the last 20 years due to globalization. This is manifested by the increasing number of surveys that are 3MC and comparative in nature. Comparisons are made in many areas including for instance official statistics, assessment of skills, and social, opinion, and market research. The ambitions and research cultures vary greatly between surveys. Even though the lifecycle of a comparative survey is quite elaborate with many process steps, we notice that many organizations in charge of 3MC surveys seem keen on covering as many populations (often countries) as possible, which leaves less room for handling all these steps. In extreme cases only a source questionnaire is developed and survey organizations in participating countries are asked to conduct remaining steps in the lifecycle with little or no guidance. There is a great risk that this approach generates estimates that are not comparable and it is important to inform stakeholders about this problem.

At the other end of the spectrum we have surveys such as the European Social Survey, the World Mental Health Survey, the Health, Ageing and Retirement Survey, and the Program for the International Assessment of Adult Competencies. These and some other surveys have strong central teams leading the efforts and assisting countries using a set of process requirements and follow-up procedures. Site visits and other meetings are also common. The idea of a strong central team has gradually evolved during the last decades. Previously it was often assumed that countries were able to follow instructions without much guidance or explanations. In the aftermath of the 1994 International Adult Literacy Survey it became obvious that this assumption was overly optimistic. It turned out that many different circumstances had made participating survey organizations deviate from prescribed implementation instructions. Since then the idea of a strong survey infrastructure and central leadership has been refined. The challenge is to sell this idea widely and to give examples of efficient infrastructures and their cost-benefits. In this chapter we give some examples.

The user situation needs clarification. Often there are conflicting national and comparative interests, which is confusing. Important decisions are made regarding policies but perhaps more on national than on international levels. The comparative aspects are, with the exception of official statistics, often dominated by league tables, when in fact the real benefits for a nation would be to investigate subgroups across nations and analyze causes of

differences. It seems as if the outreach of 3MC survey results should be more extensive and the results discussed in more detail by the public and the decision-makers. PISA, the assessment of 15-year old students using psychometrics, has succeeded with its outreach, even though the league table aspect dominates media reporting. The distance between users and producers are typically longer in 3MC surveys compared to one population or mono surveys. This distance ought to shrink by improved reporting of results and improved analyses. Researchers are often important users and they use the data to develop new theories and methods in social science, sometimes not taking limitations in the data into account.

The planning and implementation of comparative surveys is a huge undertaking. All problems experienced in a mono setting are magnified and new problems are added. The difficulties associated with developing concepts, questions and psychometric items that convey the same meaning across cultures and nations tend to be underrated but absolutely crucial to comparability. For instance, translation of survey materials is not an easy task and common perceptions that word for word translation and back translation are good for comparability are lingering. The quality issues in comparative surveys are indeed complicated. First, the various design steps are associated with error risks that vary across countries. Second, the risk perception and the management of risks also differ across countries. For instance, some error sources might not be considered due to a belief that they do not have a serious impact on comparability. Also there might not be enough resources to handle the error sources even though they are known to be problematic. Various models for capacity building will be discussed in the chapter.

A quality assurance system must be in place, which describes the requirements, justifications for their use, and a quality control system that checks that requirements are adhered to and that production is free from unnecessary variation. Few 3MC surveys have all that in place. Those who have tend to get quality control information too late to be able to intervene and rectify problems in a timely fashion. In the future we must strive for almost real-time quality control. In this work theory and methods for statistical process control must be used so that variability patterns in inflow, nonresponse, interviewer behavior, and response patterns can be diagnosed via control charts displayed on country and "global" dashboards. Technology for real-time monitoring is used in other fields such as flight tracking and could be applied in 3MC survey monitoring as well. We attempt to describe some possible future routes to develop and implement more timely quality assurance and quality control systems. This includes a discussion of the use of paradata and adaptive designs in 3MC surveys.

Few, if any, 3MC surveys are systematically evaluated or audited. There have been a few isolated quality reviews but what is needed is a more continuing approach. One option is the ASPIRE system developed for evaluations of statistical products in mono surveys. It is a system based on a mix of assessments of quality risks and actual critical-to-quality performance indicators and the assessments are performed by external evaluators to enhance objectivity. Assessments are made using a point system that makes it possible to check if improvement has taken place from one assessment to the next. We will discuss how this might be done in a 3MC context.

Roger Jowell developed ten golden rules for comparative surveys. One rule stated that the number of populations to compare, often countries, should be kept at a reasonable number. There are examples of surveys that comprise 140 countries. It is hard to imagine that such numbers and vast diversity can generate any kind of trustworthy comparability. Jowell stated that instead one should confine cross-national research to the smallest number of nations compatible with each study's intellectual needs. We will discuss the implications of this rule and some of Jowell's other golden rules with a future perspective. We will also add some rules of our own.

To gain trust many 3MC surveys need to be more transparent. Many of them lack a proper documentation of the processes and the efforts involved in controlling and improving quality. We fear that in many cases there is not much to report. One reason might be that the surveys are so extensive that all resources go to just collecting data from many countries and very little is left for sound methodology, continuous improvement, and documentation. All 3MC surveys should provide proper documentation and some already do that in excellent ways. We suggest what a minimum documentation standard might entail.

The large costs are a major deterrent to high-quality comparative surveys. A well-designed survey with resources allocated to all major design and implementation steps will cost a lot. This is one reason why country participation might decline or that requirements are not fully met. Therefore many surveys have started to explore potential cost-savers such as mixed-mode designs. The problem is that so far this experimentation shows that comparability will suffer. Also selling out a standardized approach based on input harmonization and instead face a situation where countries have lots of freedom to use methods they prefer rather than those required is not a good strategy to reach comparability. Here we will discuss the pros and cons associated with input and output harmonization. Nevertheless, the cost situation should be carefully scrutinized, since there are always activities that can be done using fewer resources or perhaps not done at all. New technology such as GIS can be a real cost-saver when locating respondents and helping interviewers administer their work. One obvious way to reduce administrative costs is to reduce the number of countries involved, which also will decrease the cost for individual countries. Of course there are also examples where generous funding is necessary to reach the research goals. There are examples where surveys have been partially funded by private financiers. Many topics that are studied by 3MC surveys are such that they may be interesting to private sponsors, for example surveys on health and education. We will explore some issues related to future funding situations.

One other way to reduce cost might be to explore other data sources as complements to the survey itself. Depending on the survey topic it might be possible to use administrative data and big data for that purpose. We will explore these possibilities and outline a roadmap for a desirable development of the use of multiple data sources in the 3MC field. Also new or not so new methodological developments such as Bayesian inference and nonprobability sampling will be briefly discussed.