## Monday, March 18th

### Session: Translation and Language Issues in Survey Research: Detecting Problems and Suggesting Solutions

**Working Remotely with Translation Teams in Implementing TRAPD: Practical Considerations**
*Alisu Schoua-Glusberg, Research Support Services*

**Abstract:** A**s** the best practice implementation of the TRAPD model (Harkness 2002), team or committee approaches to translation are best carried out with in-person meetings where the translators, the adjudicator, and potentially others sit around the table and discuss the initial translation(s) item by item.

While this approach is ideal, there are circumstances in the 3mc context in which the actors needed at the table are not all located in the same place and the translation budget does not allow for long distance travel for team meetings.

This presentation will focus on practical aspects of how to assemble and work remotely with a translation team. RSS has recently organized and managed translations with teams in Kenya, the Phillippines, and Saudi Arabia. We will discuss every step, from translator selection, briefing, team dynamics, and quality assurance.

**Surveyspeak versus Translatability: Can Gunning Fog Index Inform Questionnaire Development?**
*Danuta Przepiórkowska, University of Warsaw*

**Abstract:** Surveyspeak (including 'scalespeak') has long been known as an issue that affects surveys, especially when it comes to instruments that are targeted at the general population. This phenomenon can be found in both the original and the translated versions of survey questionnaires. Surveyspeak will also have an effect on translatability of questionnaires, rendering the translation process difficult, if not impossible at times, especially when languages from distant linguistic groups, using different structures, are involved.

Surveyspeak, which entails 'special features of questionnaire language, as found in source and target language questionnaires' (Cross-Cultural Survey Guidelines, 2011), often means that questions and scales become 'foggy' and hard to understand for the respondents. Difficulties in understanding also entail problems in responding to surveys, thus potentially affecting comparability and validity of collected data.

Can developments in modern linguistics aid question designers and translators in constructing more understandable questions and response scales? Technology developments have turned linguists' attention to the Gunning Fox index, originally developed in 1952 by Robert Gunning as a test confirming that a text can be understood by audiences with specific levels of education.
This paper will present recent academic and practical developments related to fog index for the Polish language and attempt to offer broader reflection on the possibilities to apply it to survey questionnaires, both original and translated, as a tool to test comprehensibility, enhance understanding and, potentially, reduce the total survey error.

**A New Look at Types of Probes for Testing Translated Instruments**

*Alisu Schoua-Glusberg, Research Support Services*

**Abstract:** Cognitive testing of translated questionnaires can be considered part of translation assessment. By eliciting patterns of interpretation we can assess not only the translation quality but also the validity of the translated questions and see the extent to which they are being interpreted as the source language version designers intended it.

There are different approaches to probing in cognitive testing. Some researchers prefer a highly scripted approach with a protocol that lists a number of specific probes. They craft these probes based on their expert review of the instrument and an a priori selection of possibly problematic question formulations. Other researchers prefer to elicit a narrative that will --by itself and supplemented by spontaneous probes -- reveal how the respondent's answer to the survey question relates to the respondent reality. Thus this will show how the respondent interpreted the question and whether their response was appropriately selected to fit their reality. Both approaches, with highly scripted probes or with narrative elicitation and spontaneous probing include specific probes that are asked to further elucidate the respondent thinking.

Focusing on specifics of the cross-cultural, multi-language context, this presentation will discuss a classification of probes that focuses on whether the probe asks about the question itself (or some of its features) or asks about the respondent's answer. While this way to classify probes is quite different from more traditional classifications, it can shed light on specific aspects of probing and on the right combination of probes to use in a comparative context when pretesting translated questions.

**Advance Translations Used to Enhance the Translatability of the Source Questionnaire in the ESS: Confirmation of the Usefulness in a Think-Aloud Study into French and German**

*Brita Dorer, GESIS-Leibniz-Institute for the Social Sciences*

**Abstract:** In multilingual surveys, questionnaire translation problems or errors are often caused by issues in the source questionnaire. For minimising such problems, the ESS has carried out, since its 5th round in 2009, systematic ‘advance translations’ in order to detect such problems before finalising the source text. For doing so, national ESS translation teams, consisting of both translators and survey researchers, carry out translations of a pre-final version of the source questionnaire, with the purpose of spotting translation problems. These comments are considered when finalising the source text. The problems pointed out range from intercultural adaptation issues to, for instance, grammatical or syntactical structures requiring complicated translations into a specific target language that may have a negative impact on the comparability between all resulting final translations when fielded. Changes in the source text triggered by advance translation range from rephrasing source text elements to adding footnotes to explain ambiguous source text terms.

The author tested the usefulness of this method in a series of think-loud tests: experienced questionnaire translators translated 22 items – in their version before and after the advance translation – into French and German. The think-aloud protocols were analysed both qualitatively and quantitatively. The usefulness of advance translation for enhancing the translatability of the source text was confirmed in this think-aloud study.
This paper describes the method of advance translations as applied in the ESS, some typical source questionnaire issues detected, as well as the think-aloud study, its analysis and results.

Conclusions will be drawn on questions like: does the success of advance translation depend on the source text problems detected or on the changes made because of advance translation? Which role does the interplay of languages of advance translation and think-aloud study play? Is think-aloud a valid method for evaluating advance translation?

**Mechanisms of Close versus Free Questionnaire Translation: Qualitative Findings from an Experiment in Estonian and Slovene**

*Brita Dorer, GESIS-Leibniz-Institute for the Social Sciences*
Lydia Repke, GESIS-Leibniz Institute for the Social Sciences

**Abstract:** To what extent do different translation approaches (i.e., close and adaptive) shape the data we collect within multinational survey projects? Are there specific topics or linguistic patterns where translators feel the need to adapt? Triggered by the common belief that close translation yields more comparable data than adaptive translation, this paper systematically examines different translation versions and the resulting responses of a translation experiment conducted within the cluster project SERISS. For this experiment, three translation teams into Estonian and three into Slovene were instructed to translate 60 English source items by applying both competing translation approaches. The source and the translated questionnaires were fielded in the CROss-National Online Survey (CRONOS), Wave 5, with participants from Great Britain, Estonia, and Slovenia.

For the analysis, native speakers of both target languages first assessed whether the translations were adaptive or close and provided explanatory back-translations into English. Next, they evaluated the overall translation potential of each source item (i.e., the theoretical translation space with all possible and meaningful translations) on a 5-point Likert scale ranging from -2 (close) to 2 (adaptive). Finally, they assessed the translation score of each translation (i.e., the realized translation) on a 7-point Likert scale ranging from -3 (overly close) to 3 (overly adaptive). They did this for each question and answer scale separately. Based on this information, we analysed the occurring translation patterns and combined them with the responses.

Our preliminary analysis shows that it is not always possible to apply both approaches for all items. This speaks to the importance of not having a “one-fits-all” translation strategy. For example, not all items can be translated closely. Instructing translators to do so anyway may lead to bad or wrong translations. We are currently developing recommendations on how to handle adaptive translation approaches in multilingual surveys.

## Session: Comparability of Measures

**The Total Survey Error Paradigm and Comparison Error: A Component-Level Evaluation**

*Tom W. Smith, NORC at the University of Chicago*

**Abstract:** Earlier work has demonstrated that the total survey error (TSE) paradigm can and should be applied to cross-national and cross-cultural studies. Both for each component of TSE and overall across all components, a difference between one survey and another survey creates a comparison error that reduces comparability and thus undermines the substantive validity of comparative research. In the basic TSE framework proposed by Smith, there are 35 components. While the first goal would be to eliminate error in each component in each survey, this is impractical for most components and impossible for others (e.g. there will also be random sampling error whenever a sample is utilized). For

comparative surveys the twin practical goals are a) to identify and minimize error across components in general and b) to reduce differential error in components across surveys (e.g. across countries). While this general goal exists across all components, how it can be achieved varies greatly across components. In some cases comparison error can and should be minimize by doing things "exactly" the same across surveys. In other situations things cannot be done the same way across surveys, but can be done in an equivalent manner. In still other instances, neither exactitude nor equivalence can be achieved and comparison error must be managed and adjusted for rather than eliminated or minimized. Why these different approaches need to be used for different components is illustrated for several components including sampling, non-response, data collector, question wording, and data processing.

**Regular o Pasable: Improving Measurement Properties of Self-Rated Health for U.S. Latinos through Alternative Translation**

*Sunghee Lee, University of Michigan*
*Fernanda Alvarado Leiton, University of Michigan*
*Elizabeth Vasquez, State University of New York, Albany*
*Rachel Davis, University of South Carolina*

**Abstract:** Self-rated health (SRH) is a powerful indicator of health. However, in the U.S., Latinos are found to report consistently less favorable SRH than non-Latino Whites, ceteris paribus, which may be attributed to measurement error related to Spanish translation of the SRH question. In particular, the connotation comparability of regular, a Spanish translation of the SRH response category, "fair," has been questioned.

We translated "fair" into two Spanish versions: regular and pasable. A split-half experiment was implemented in three studies that included a total of 3,264 Latino adults and 738 non-Latino White adults in the U.S. In the experiment, Latino respondents interviewed in Spanish were randomly assigned to be asked one of the two Spanish versions of SRH.

Spanish-interviewed Latino respondents reported substantively more favorable SRH when using pasable than regular. When using SRH with pasable instead of regular, larger difference between respondents with positive versus negative SRH were also observed on the frequency of doctor's visits and feeling younger than the actual age. Unlike the SRH version with regular, Latino-White disparities were attenuated with pasable when accounting for correlates of SRH.

Conclusions. We recommend using pasable instead of regular in Spanish translations to improve equivalence in meaning in cross-lingual and cross-cultural measurement of SRH.

**Measuring School Children's Attitudes towards Minorities in Poland and Switzerland**

*Charlotte Clara Becker, University of Cologne*
*Eldad Davidov, University of Cologne*
*Jan Cieciuch, Cardinal Stefan Wyszyński University in Warsaw*
*René Algesheimer, University of Zurich*
*Martin Kindschi, University of Zurich*

**Abstract:** In the last decade there has been a shift to the far right in many European Societies. This is especially true for attitudes towards minorities and the public debate of such. For years, social scientists have been interested in these attitudes and their development over time. Here the focus usually laid on showcasing the opinions of native adults. In contrast, little is known about children's and teenager's attitudes towards minorities. However, in order to gain insight into societies sentiments as a whole, it is essential not to overlook its youngest members.

This lack of research on children might be due to the previously used measurements, which were unstandardized, costly and time-intensive, leading to small, regional samples. We aim to overcome these problems, by introducing and evaluating a new, child friendly, easily administrable and universally applicable way of measuring children's attitudes towards minorities: Picture based, self-administered questionnaires.

For the analyses we use a Polish-Swiss panel data set (2015-2017) collected among school children by the research priority program "social networks" at the University of Zurich. The sample includes 5332 children aged 8 to 19, separated in three age groups, 4th graders, 7th graders and 9th (Switzerland)/10th (Poland) graders. The questionnaire includes pictures of four different minorities (blind, in wheelchair, black and Muslim). For each minority the same four questions are asked. These questions are then used as the items in multigroup confirmatory factor analyses.

Preliminary results show that the new picture-based measurement works well. Besides having good model fit, the measure also holds when checking for external validation. Further, invariance testing revealed that the measurement can not only be used across different age groups equivalently, it also achieved metric, in some cases even scalar invariance across the two countries. This shows that measurements comparability in children can be achieved with picture-based questionnaires.

**Household and Personal Income Measures in Cross-National Survey Projects: Assumptions, Usability and Comparability**

*Marta Kołczyńska, Institute of Philosophy and Sociology, Polish Academy of Sciences*
*Ilona Wysmułek, Institute of Philosophy and Sociology, Polish Academy of Sciences*
*Denys Lavryk, Institute of Philosophy and Sociology, Polish Academy of Sciences*

**Abstract:** Individual economic status is an important element of many social theories, whether it is understood as an indicator of social position, or in terms of returns on investment in, for example, education. To provide measures of economic status, social surveys often include questions about respondents' income, either personal or of the whole household, or both. Empirical studies with survey data frequently use income measures, often without justifying why the given item is appropriate. To justify the use of one income measure over the other is challenging, as the question of the quality, underlying assumptions, usability and comparability of different income measures are rarely discussed explicitly.

In this presentation we analyze the availability and diversity of personal and household income measures in 1721 national surveys from 22 cross-national survey projects around the world in the period 1966-2013, including the World Values Survey and European Value Study, the European Social Survey, the International Social Survey Programme, and others, selected for ex-post harmonization in the Survey Data Harmonization project (SDR v.1 dataset, dataharmonization.org). Of all national surveys that we investigated, a household income measure is available in 1177, and personal income in 453 national surveys. 419 surveys contain both items, which allows us to analyze the association between household and personal income measures, and factors that might influence it.

We find that, indeed, the correlation between household and personal income tends to be positive and strong among men, but less so among women. The strength of the association between the two income variables also varies across age groups and education levels. These findings suggest caution when using household income where the theory implies personal income, and vice versa. We conclude by discussing ways of harmonizing income variables from cross-national surveys in light of their applications in substantive analyses.

## Session: Interviewer Variations & Training

### Measuring Interviewer Compliance with Regard to Question Deviations in a Multi-Language Survey in Zambia

*P. Linh Nguyen, University of Essex; University of Mannheim*

**Abstract:** International development projects are usually evaluated on their impact by using survey data as evidence. Due to limited outreach of telephone and internet devices and infrastructure, interviewer-administered face-to-face (F2F) surveys are and will remain the principal data collection tool in developing countries. Although survey data is frequently collected, the common practices for questionnaire pre-testing, as well as systematic evaluation of the interview administration process, have yet to be established broadly among practitioners in non-Western countries. In this light, this study presents new empirical insights about the quality of survey data collected in a Zambian setting. The existence of multiple ethnicities and multiple languages in Zambia, like in the majority of African countries, pose challenges to any data collection as the interviewer is multilingual and thus, able and sometimes asked to do on-the-spot translation in a different language. The analysis draws on data from a face-to-face survey on standards of living, economic situation and financial behavior in rural or semi-urban areas of Zambia. It was conducted in 2016 with more than 2000 members of selected collective savings groups who are beneficiaries of a development program. The focus of this investigation is to explore to what extent interviewers comply in delivering five selected factual and attitudinal questions in a standardized manner given the multi-lingual context. The questionnaire was translated from English into the respective dominant local language of the three survey locations. About 4,000 interviewer-respondent-interactions on those questions were coded following a behavior coding scheme to study whether questions were administered following the pure standardized interviewing approach or whether there were minor or major deviations. The results from the behavioral coding, as well as methodological issues in the development of the coding scheme and intercoder reliability, shall provide a base for future improvements in interviewer training adjustment or questionnaire revision.

**Development of Bilingual Interviewer Training at the U.S. Census Bureau**

*Patricia Goerman, U.S. Census Bureau*
*Mikelyn Meyers, U.S. Census Bureau*
*Yazmin Garcia Trejo, U.S. Census Bureau*

**Abstract:** Over the last two decades, researchers at the U.S. Census Bureau have focused on the development of messages to encourage survey participation at the doorstep in multiple languages. The research has included ethnographic observation of face-to-face interviews in multiple languages, expert review of translations of messages to tailor them to different language and cultural groups, and focus groups to get respondent feedback on draft messages in English, Spanish, Chinese, Korean, Vietnamese, Russian and Arabic. This research has provided a wealth of information and recommendations on tailored messages for use at the doorstep. Our next step is to get these recommendations into the field and provide support for our bilingual field interviewers. The work focuses on transforming past findings and recommendations into concrete training materials. This talk will give an overview of our implementation plans including: 1. A review of current Census Bureau training materials that include language and cross-cultural topics, 2. Plans for the design of two field interviewer training modules for use in the 2020 Census, and 3. Design of a new 2020 observation study and a field experiment related to bilingual interviewer training. Finally, this talk seeks to generate discussion of cross-cultural and/or bilingual interviewer training that other workshop participants have at their organizations.

**Why do Interviewers Vary on Interview Privacy and Does Privacy Matter?**

*Zeina Mneimneh, University of Michigan*
*Julie de Jong, University of Michigan*
*Jennifer Kelley, University of Michigan; University of Essex*

**Abstract:** The presence of a third person 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). Interviewers play an essential role in requesting, achieving, and reporting on the private setting of the interview. Our recent work has shown that the rate of interview privacy varies significantly across interviewers; while some interviewers report high rates of privacy among their interviews, others report low rates of privacy for the interviews they administered (Mneimneh et al., 2018). Yet, there is a lack of understanding of what explains such interviewer variation in interview privacy. Do certain interviewer characteristics such as experience, socio-demographics, and attitudes towards privacy explain such variations? What about the measurement quality of the privacy observation measures interviewers collect? Is it possible that section-specific measures (where the interviewer collects such observations right after questionnaire sections) show less interviewer variation than end-of-the-interview measures (the commonly used method of collecting interview privacy data) because of potential differential recall across interviewers?

This paper explores these research questions for the first time using data from a national mental health survey conducted in the Kingdom of Saudi Arabia. A total of 4000 face-to-face interviews were completed using a computer assisted personal interviewing (CAPI) mode. Interviewers were 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. We use these two types of observations and measure the contribution of interviewer variation to these estimates. We then compare predictors of

interview privacy for each of the two types of observations using a series of multilevel models focusing on the effect of interviewer-level characteristics (while controlling for respondent and household level characteristics). Findings from this paper will have important practical implications related to training interviewers on requesting, maintaining, and reporting information on the private setting of the interview.

## Session: Achieving Comparability and Translational Equivalence in Cross-Cultural Survey Research Using Technology from Natural Language Processing

### Complementing Comparative Surveys Through the Use of Word Embeddings

*Magnus Sahlgren, Research Institutes of Sweden*

**Abstract:** The field of survey-based country comparative research has traditionally solely relied on survey questionnaires administered to controlled populations. In recent years, several problems with this methodology has been identified and discussed. One major obstacle is the increasing difficulty and cost of finding sufficient amounts of representative respondents in the various languages. Decreasing response rates are not only devastating for the representativeness of the surveys in relation to a defined population; it has also forced researchers to opt for minimizing the length of the actual survey questionnaires, something which has made it more difficult to measure more complex concepts that ideally demands longer querying batteries.

This paper explores a complementary approach to country comparative surveys that uses Natural Language Processing techniques applied to unsolicited data harvested from social (and other) media on the Internet. Such data have the advantages of being contemporary, freely available, in vast amounts, and in many, if not most, languages covered in traditional country comparative studies. We use the web data to build representations of word usage patterns in continuous vector space models (so-called word embeddings). The word representations can be used to compile data-driven thesauri, which rank semantically related terms. By analyzing the semantic neighbors of a term, we can get a good understanding of how the term is being used in the data.

We suggest that such analyses can act as a complement to traditional survey methods in cases where the question concerns how citizens conceive certain concepts, such as happiness, corruption, or the meaning of democracy, all concepts that are as multifaceted as they are time-consuming to measure in traditional ways. The research question in such comparative studies is often cast as a categorization problem, in which the task is to determine which of a predefined number of categories are more salient for citizens of different countries. As an example, a survey question about the satisfaction with democracy can concern different categories at different levels of abstraction across countries such as performance, principles, or actors. By categorizing the semantic neighbors of terms that represent the concepts in question (e.g. the term ‘democracy’) with respect to these different categories, we arrive at a measurement of how prevalent these categories are in unsolicited language use in the various languages. The alternative in a traditional survey manner would instead be to rely on a large predefined battery of survey-items, which is empirically limited and costly. An attractive byproduct of such analyses is that the categorized data can be used as training data for machine learning models that can automate the classification, which enables both large-scale and extremely resource-effective surveys.
This paper introduces the general methodology, and provides examples of studies using survey items such as the satisfaction with the working of democracy, taken from the Comparative Studies of Electoral Systems (CSES) dataset module 3 and 4, Corruption taken from the Transparency International and

happiness from the World Value Survey wave 6. From these survey-based studies we are able to rank countries according to their average levels of democratic satisfaction, presence of corruption and happiness but we still do not know what these concepts actually means to people. By combining survey data with online language data, we are able to capture both conceptual meaning as well as overall differences in levels.

We also discuss challenges with the proposed approach, such as how to find relevant, comparable, and representative online data, how to account for typological differences when building word embeddings for different languages, and how to interpret results from nearest neighbor analysis.

**The Meaning of Democracy – Using a Distributional Semantic Lexicon to Collect Co-Occurrence Information from Online Data across Languages**

*Sofia Axelsson, University of Gothenburg*
Stefan Dahlberg, University of Gothenburg

**Abstract:** The literature on public support for democracy has revealed significant cross-country differences in people's attitudes towards democracy. Explanations for such variations can be found in the survey literature on "diffuse" versus "specific" support for democracy; whereas the former refers to support for the democratic principles in a more abstract sense, and is generally found in consolidated democracies, the latter concerns more specific support for political performance of democracies, which is more prevalent in new democracies (see Easton 1975; Norris 1999; Dahlberg & Holmberg 2012; Linde & Ekman 2003). Yet, how are we to know what democracy means to the people answering surveys, and thus be able to identify what they are expressing support for? Some scholars have further disentangled survey batteries in order to capture different notions of democracy among citizens living under different cultural and institutional settings, arguing that the concept of democracy becomes distorted in authoritarian settings (Welzel & Kirch 2017; Welzel 2013). However, the literature lacks systematic comparative studies outside the realm of surveys that takes institutional, cultural, and, importantly, linguistic variations into account.

Cross-cultural survey research rests upon the assumption that if survey features are kept constant to the maximum extent, data will remain comparable across languages, cultures and countries (Diamond 2010). Yet translating concepts across languages, cultures and political contexts is complicated by linguistic, cultural, normative or institutional discrepancies. Recognizing that language, culture and other social and political aspects affect survey results has been equated with "giving up on comparative research", and, consequently, the most commonly used "solution" to equivalence problems has been for researchers to simply ignore the issue of comparability across languages, cultures and countries (Hoffmeyer-Zlotnik & Harkness 2005; King et al. 2004).

This paper contributes to the debate, using a distributional semantic lexicon, which is a statistical model for collecting co-occurrence information from large text data (Turney & Pantel 2010). The lexicon represents terms as vectors in multi-dimensional context space, where relative similarity between vectors indicate similarity of usage, which is often equated with semantic similarity. The method is motivated by a structuralist meaning theory known as the "distributional hypothesis", which states that words with similar meanings tend to occur in similar contexts, and that the contexts shape and define the meanings of the words (Sahlgren 2006). Compared to other methodological approaches aimed at identifying and measuring cross-cultural discrepancies, this approach has the advantage of enabling us to analyze how the concept of democracy is used in its "natural habitat" (Wittgenstein 1958). Collecting

geo-tagged language data from editorial and social online media thus allows us to explore the varieties in understandings of democracy across different languages and countries, and to map the ways in which democracy is used among populations and societies worldwide, also across different institutional settings and regime types.

**Lost in Translation – How Differences in Word Intensity Affect Citizens' Satisfaction with the Working of Democracy**

*Stefan Dahlberg, University of Gothenburg*
*Magnus Sahlberg, Research Institutes of Sweden*

**Abstract:** Cross-cultural survey research rests on the assumption that if survey features are kept constant to the maximum extent, data will remain comparable across languages, cultures and countries (Smith 2003). Yet translating concepts across different settings is complicated by linguistic, cultural, normative as well as political and institutional discrepancies (Harkness 1999). The research community is not unaware of this and survey methodologists have long wrestled to solve the issues of cross-cultural comparability and semantic equivalence in country comparative surveys. Standard practices have been to work with careful question wording and response scale design as well as with translational procedures (including back-translation, pre-testing and TRAPD methods).

Even as general survey methodology by now constitutes a vast field of research, less effort has been devoted to the issue of cross-country comparability. With respect to variations in response patterns caused by translational effects, most studies have focused on within-country comparisons. Still, we know that language effects do exist and impact on cross-cultural variations in response patterns (Zavala-Rojas 2018). For example, response-category differential item functioning (DIF), which occurs when different groups of respondents interpret the response labels of an ordinal scale differently, is particularly problematic in survey questions with ordinal response categories (Villar 2009). One technique specifically developed to mitigate the problem of DIF is to include anchoring vignettes in the survey, a method introduced by King et al. (2004). Anchoring vignettes do not, however, solve the problems associated with DIF in cross-cultural surveys that have already been conducted.

Recognizing that language and culture coupled with other socio-political aspects affect survey results has sometimes been equated with "giving up on comparative research" and, consequently, the most commonly used "solution" to equivalence problems has been for researchers to simply ignore the issue of comparability across languages, cultures and countries (Hoffmeyer-Zlotnik & Harkness 2005; Wierzbicka 2004). However, recent developments in the field of natural language processing (NLP) has provided social scientific scholars with innovative ways and means to address inconsistencies in comparative surveys, which ultimately can be geared toward the issue of comparability and measurement equivalence.

This session focuses on the possibilities of applying technology from NLP to survey research in general and to survey translation in particular. More specifically, it aims at exploring NLP techniques vis-á-vis standard survey practices to identify, counter and control for the effects of language and culture in comparative surveys. The session further aims at exploring applications of NLP to online data to capture sentiment and meaning differences in conceptual issues across cultures as a possible way to enrich and complement traditional survey methodology and survey data.

**Measuring Issue Ownership using Word Embeddings**

*Amaru Cuba Gyllensten, Research Institutes of Sweden*
*Magnus Sahlgren, Research Institutes of Sweden*

**Abstract:** Social Media Monitoring (SMM; i.e. monitoring of online discussions in social media) has become an established application domain with a large body of scientific literature, and considerable commercial interest. The subfields of Topic Detection and Tracking (Allan et al., 1998; Sridhar, 2015) and Sentiment Analysis (Turney, 2002; Pang and Lee, 2008; Liu, 2012; Pozzi et al., 2016) are both scientific topics spawned entirely within the SMM domain. In its most basic form, SMM entails nothing more than counting occurrences of terms in data; producing frequency lists of commonly used vocabulary, and matching of term sets related to various topics and sentiments. More sophisticated approaches use various forms of probabilistic topic detection (such as Latent Dirichlet Allocation) and sentiment analysis based on supervised machine learning.

The central questions SMM seeks to answer are "what do users talk about?" and "how do they feel about it?". Answers to these questions may provide useful insight for market research and communications departments. It is apparent how product and service companies may use such analysis to gain an understanding of their target audience. It is also apparent how such analysis may be used in the context of elections for providing an indication of citizens' opinions as manifested in what they write in social media. There are numerous studies attempting to use various forms of social media monitoring techniques to predict the outcome of elections, with varying success (Bermingham and Smeaton, 2011; Ceron et al., 2015).

Most notably, the recent examples of the inadequacy of standard opinion measuring techniques to forecast the most recent US election and the Brexit demonstrate that for certain questions related to measuring mass opinion, standard SMM techniques may be inadequate. Political scientists have used the concepts of agenda setting and issue ownership to explain voter choice and election outcomes (Kl¨uver and naki Sagarzazu, 2016; Kiousis et al., 2015; Stubager, 2018). In short, the issue ownership theory of voting states that voters identify the most credible party proponent of a particular issue and cast their ballots for that issue owner (B´elanger and Meguid, 2008). Agenda setting refers to the media's role in influencing the importance of issues in the public agenda (Mccombs and Reynolds, 2002). Note that current social media monitoring techniques are unable to measure these concepts in a satisfactory manner; it does not suffice to measure the occurrence of certain keywords, since most parties tend to use the same vocabulary to discuss issues, and sentiment analysis does not touch upon the issue ownership and agenda setting questions. What is needed for measuring issue ownership and agenda setting is a way to measure language use, i.e. when talking about an issue, to which extent does the language used align with issue owner A vs. issue owner B.

We argue that issue alignment can be seen as a kind of semantic source similarity of the kind "how similar is source A to issue owner P, when talking about issue X", and as such can be measured using word/document embedding techniques. To measure that kind of conditioned similarity we introduce a new notion of similarity for predictive word embeddings. This method enables us to manipulate the similarity measure by weighting the set of entities we account for in the predictive scoring function. The proposed method is applied to measure similarity between party programs and various subsets of online text sources, conditioned on bloc specific issues. The results indicate that this conditioning disentangles similarity.

We can, for example, observe that while the Left Party representation is, overall, similar to that of nativist media, it differs significantly on nativist issue, while this effect is not seen to the same extent on more mainstream left wing or right-wing media.

**Do Not Take Online-Mediated Text for Granted: Heuristics for Assessing Limitations of Representativity in Online Text Data**

*Jonas Andersson Schwarz, Södertörn University*
*Fredrik Olsson, Research Institutes of Sweden*
*Sofia Axelsson, University of Gothenburg*

**Abstract:** This paper reflects an ongoing project intended to validate the use of online text data as a complement to traditional surveys and polls. In order to assess the comparability of online-mediated text and survey-generated text – where natural language processing (NLP) is employed to extract patterns in each respective corpus – the nature of the actual text sources for each corpus needs to be addressed, in terms of validity, reliability and, in particular, representativity.

The problem of representativity of online text data has been covered by numerous scholars (e.g. Andersson Schwarz & Hammarlund 2016; Blank 2017; Malik & Pfeffer 2016; Mellon & Prosser 2017; Ruths & Pfeffer 2014; Tufekci 2016) and this paper aims to add to this growing body of literature whilst simultaneously offering a useful step-by-step heuristic for assessing the usefulness online text data for social scientific research purposes.

It is established that online-mediated text has certain representation biases and/or limitations, not only in terms of demography (certain age groups and media literacy groups are more represented than others) but also in cognitive terms (certain data is registered, or even registrable through interfaces while other data is not), rhetorical and normative terms (certain expressions, conceptions and discourses are favored with particular normative slant), as well as social networking (certain in-groups are generated, and network typology determines the communication across groups) and political and economic incentives of various platform owners (where certain modes of interaction and engagement are favored due to e.g. advertising business models).

Where the issue of representativity is concerned, there are methodological parallels with traditional mass media content analysis, which has grappled with similar challenges for decades. No one expects newspaper prose to be directly representative of public opinion, yet it is still used for quantitative assessments of discourse and can be highly useful for analyses of social scientific problems. The key, we argue, is to properly contextualize the findings from online text data and, if possible, find suitable quantitative techniques for mitigating known biases. Importantly, in order to fully utilize online text data in a comparative setting – e.g. to complement and improve traditional survey data – online data sources must cover different languages and countries and be comparable across different cultural contexts.
As a means to improve such contextualization, this paper offers a heuristic for assessing the quality and scope of different sources of online text data. More specifically, through a step-by-step framework, the following dimensions are addressed: a) quantitative measures of language distribution (i.e. how well do providers of online text data fare in the coverage of different languages?); b) data provenance (i.e. what qualities define different data sources and what indicators are suitable for assessing the quality of sources?); c) incomplete data samples versus complete data samples (i.e. what data size is sufficient for the application of various NLP methods and what are the best sampling practices?); and d) quantitative measures of intra-linguistic patterns (i.e. what indicators are suitable for rapid assessment of text samples and how are intra-linguistic patterns addressed?).

# 2018 CSDI Workshop Abstracts

## Tuesday, March 19th

### Session: Developing Interactive Tools for Cross-National Surveys: Results from the SERISS Project

**Introducing an Electronic Fieldwork Management System in the European Social Survey**
*Sarah Butt, City, University of London*

**Abstract:** A key challenge facing all social surveys is to monitor and manage fieldwork effectively. Cross-national surveys, such as the European Social Survey (ESS, www.europeansocialsurvey.org) face particular challenges as they try to monitor fieldwork conducted by survey agencies in different countries using different methods and systems to manage and monitor data collection. This can result in delays and inconsistencies in the flow of information and even loss of information between the fieldwork agencies, interviewers in the field and the central survey team, making it difficult to monitor fieldwork in an effective, consistent and timely manner across countries.

The ESS is investing in a new electronic Fieldwork Management System (FMS) with the aim of providing all stakeholders with access to consistent and timely data on fieldwork progress across countries throughout the fieldwork period. The tool has been developed with CentERdata as part of the SERISS project (www.seriss.eu). Two versions of the FMS are available: In the first, countries have access to a mobile app which interviewers use to manage their cases and collect contact data and which syncs automatically with an online database accessible to the national survey agency and ESS central fieldwork team. In the second, agencies wishing to collect contact data using their own in-house system can alternatively upload weekly monitoring information to the FMS online portal using a pre-defined template. Both versions of the tool are currently in use in the field for ESS Round 9.

This presentation provides an overview of the FMS. It discusses some of the methodological, practical and technological challenges associated with rolling out the tool in 25+ countries. It also demonstrates how access to standardised real time fieldwork monitoring data as provided by the FMS is benefiting the ESS Fieldwork Team and better informing our understanding of fieldwork flows in different European countries.

**Documenting the European Social Survey Questionnaire Design Process Online Using the Data Documentation Initiative (DDI)**
*Hilde Orten, NSD - Norwegian Centre for Research Data*
Stig Norland, NSD - Norwegian Centre for Research Data

**Abstract:** The Questionnaire Design and Documentation Tool (QDDT) is a web-based tool developed to assist large-scale survey teams in developing thematic questionnaire modules for their survey.
The tool is designed to capture the development of concepts and questions throughout the questionnaire design process, to reuse question-related metadata components over time, to version them, and to publish content at specific milestones.

The metadata standard DDI-Lifecycle version 3.2 is used as a basis for the conceptual model of the tool, which additionally facilitates reuse of metadata and interoperability between applications.

The complex questionnaire design process of the European Social Survey (ESS) is use case for the QDDT, and the tool is currently used by the ESS for the questionnaire module design in its 10th round.
The presentation focuses on how DDI is used to facilitate the needs of the ESS to keep track of and document changes to concepts and questions over time, to reuse questions and response domains, and to be able to export content to other tools.

**Challenges and Best Practices When Designing a Mobile Application for a Probability-Based Internet Panel**

*Geneviève Michaud, Sciences Po Paris*

**Abstract:** Launched in 2012, ELIPSS (Longitudinal Internet Studies for Social Sciences) is the first French online representative panel dedicated to the social sciences research community in 2012. Each panel member is provided with a touch screen tablet and a 4G connection. The project has now hit its stride with around 2500 panel members and 70 surveys conducted so far.
The Center for Socio-Political Data (CDSP) is responsible for the implementation of this project within the DIME-SHS (Data, infrastructures and survey methods in the humanities and social sciences) consortium.

The CDSP IT team designs the software tools that collect data and paradata from the ELIPSS panel. These tools range from web services facilitating fieldwork management to a tailor-made Android application for collecting data directly from devices.

First, designing an Android application collecting data for research purposes represents a real challenge. Very few guidelines are publicly available except for designing market-oriented mobile applications. Second, for the ELIPSS project, conflicting objectives have to be reconciled. One survey is fielded monthly and tight deadline are regularly to be met.

This presentation will provide a concrete view of the design process of a mobile application for data and paradata collection from an online panel. We'll also provide a feedback on how to collaborate with a device manufacturer, how to deal with confidentiality concerns in data and paradata collection, how to address ergonomy issues when designing questionnaires, and generally how to address complex research questions.

**The Question Variable Database (QVDB): A Portal and Documentation Tool for the ESS**

*Benjamin Beuster, NSD - Norwegian Centre for Research Data*

**Abstract:** The Question Variable Database (QVDB) with the Colectica platform as technical backbone can be described as a syst **Documenting the European Social Survey** em for storage and retrieval of questions and variables, and facilitating reuse of their metadata and metadata components. The overall aim of the QVDB is to serve the ESS in their business processes of specifying, documenting, versioning and disseminating survey data.

The European Social Survey (ESS) is an academically driven cross-national survey that has been conducted across Europe since its establishment in 2001. Every two years, face-to-face interviews are conducted with newly selected, cross-sectional samples.
In particular, this presentation describes how the QVDB was populated with 8 waves of ESS data. Then will give examples on how variables from different points in time correspond to each other (variable

harmonization) and we demonstrate how we use the tool for data production and data documentation processes.

There are other tools compatible with the QVDB: the SERISS Questionnaire Design and Development Tool (QDDT) to document and retrieve information on the design process of developing a cross-national survey questionnaire, and the CESSDA Euro Question Bank which provides a central search facility across all CESSDA's survey questions in different languages. The aim for these tools is to be able to exchange metadata so that metadata collected at one stage of the survey life cycle can be reused in another stage of the survey life cycle.

**The myEVS Project Management Portal: A Case Study for Survey Projects**

*Evelyn Brislinger, GESIS-Leibniz-Institute for the Social Sciences*
*Dafina Kurti, GESIS-Leibniz-Institute for the Social Sciences*
*Masoud Davari, GESIS-Leibniz-Institute for the Social Sciences*
*Markus Quandt, GESIS-Leibniz-Institute for the Social Sciences*
*Claus-Peter Klas, GESIS-Leibniz-Institute for the Social Sciences*

**Abstract:** In particular, larger and cross-national survey projects are faced with the question of how they can simplify the planning of the survey, support collaboration between distributed actors, streamline workflows, and provide all participants with insight into on-going processes. With those questions in mind, a SERISS-funded team of researchers from GESIS, FORS, and the University of Tilburg has adapted the open source version of a commercial collaboration system (eXo Platform®) to the organization and workflow structure of the current European Values Study wave 2017. The myEVS Portal was opened for general use by all EVS partners in October 2017 and is since that time the exclusive internal communication means for all operative and most strategic purposes. It offers guidance by setting up standardized step-by-step processes and document storage structures. So called "workspaces" designated for national teams and central planning groups make sure that the information flows are targeted towards the roughly 140 portal users from 44 countries across Europe. With the first release of the EVS 2017 integrated dataset, the first national teams have gone through all survey lifecycle phases on the portal. This allows us to present initial experiences from the perspective of the different actors, the central planning groups and the teams from the participating countries.

## Session: Mode Preferences & Differences

**Possibilities of Using Declarative and Behavioral Data in Predicting Respondents Survey Mode Preference in Poland**

*Adam Rybak, Institute of Sociology, University of Adam Mickiewicz, Poznan*

**Abstract:** The constant decrease of response rates is one of the main challenges faced by the survey research. One of the possible strategies for coping with this problem is the mixed-mode survey design. The key concept allowing to better understand the possibilities and difficulties connected to mixed-mode research is the mode preference hypothesis. It states, that respondents have (quasi-)constant preference toward one mode of survey compared to another. It does not specify whether these respondents are fully aware of their real preference or not. Also, the differentiation of effects of preference on contacting, and obtaining cooperation from respondents is not well established.

This presentation is based on the data from two sources: Mixed-mode survey experiment parallel to European Social Survey Round 7 in Poland and the additional set of questions used in ESS8 in Poland. The first set of data contains information about behavioral preferences – respondents faced a choice of modes. Nature of second is declarative – all respondents were asked about preferred mode during the face-to-face survey. There are also two aims: Identifying possible predictors of mode preference and analyzing the usefulness of declarative data. Due to more practical, than theoretical approach predictors analyzed are characteristics of respondents present in PESEL survey frame.

Both datasets combined are used to create logistic regression models controlling the character of data. The result of the analysis could be used in designing the new mode-mixing survey design for use in Poland.

**PC versus Mobile Survey: Are People's Life Evaluations Comparable?**

*Francesco Sarracino, STATEC*
*Cesare Fabio Antonio Riillo, STATEC*
*Malorzata Mikucka, MZES*

**Abstract:** The literature on mixed mode surveys has longly investigated whether face-to-face, telephone, and online survey modes permit to collect reliable data. Less is known about the potential bias associated to using different devices to answer online surveys. We compare subjective well-being measures collected over the web via PC and mobiles to test whether the survey device affects people's answers to subjective questions. We use unique, nationally representative data from Luxembourg which contains five measures of subjective well-being collected in 2017. The use of multinomial logit with Coarsened Exact Matching indicates that the survey tool affects life satisfaction scores. On a scale from 1 to 5, where higher scores stand for greater satisfaction, respondents using mobile phones are more likely to choose the highest well-being category, and less likely to choose the fourth category. We observe no statistical difference for what concerns the remaining three categories. We test the robustness of our findings using three alternative proxies of subjective well-being. Results indicate that survey tools do not induce any statistically significant difference in reported well-being. We discuss the potential consequences of our findings for statistical inference.

## Session: Sample Management Systems and Sampling Innovations

**Designing a Sample Management System for use in a Cross-national On-line Web Panel: Initial Thinking and Ideas**

*Gianmaria Bottoni, Euroepan Social Survey HQ, City, University of London*
*Rory Fitzgerald, European Social Survey HQ, City, University of London*
*Nicolas Sauger, Sciences Po*
*Genevieve Michaud, Sciences Po*
*Quentin Agren, Sciences Po*

**Abstract:** The ESS recently experimented with the worlds' first input harmonised probability based cross-national web panel in three countries by recruiting panel members who had taken part in the face-to-face survey. The experiment took place in Estonia, Great Britain and Slovenia (the CRONOS web panel).

A key challenge identified during the CRONOS experiments was the absence of a sample management system that was well suited for use in a multi-country environment and which could also meet data protection requirements.

This paper will describe initial plans for developing a sample management system for a cross-national web panel that meets the needs of different surveys in such a complex environment and which also links seamlessly to a survey platform. Proposals for content will be outlined with the key fields for sample management presented. In addition functionality will be discussed such as contact modes (SMS, postal and e-mal) and user accounts. Finally user profiles rights will be examined with a view to receiving feedback on the draft specification from CSDI members.

**Respondent Driven Sampling for Immigrant Populations: An Application to Foreign-Born Korean Americans**

*Sunghee Lee, University of Michigan*
*Ai Rene Ong, University of Michigan*
*Chen Chen, University of Michigan*
*Michael Elliott, University of Michigan*

**Abstract:** Changes in the racial/ethnic composition of the U.S. population, coupled with growing evidence supporting the need for data on more granulated racial/ethnic categories (e.g., Korean rather than Asian) have made increasing minority data availability a critical goal for disparity research. Achieving this goal under the traditional sampling framework is prohibitively resource intensive. Respondent driven sampling (RDS), a variant of snowball sampling, may be an alternative for due to its cost advantages. Further, strong social ties within racial/ethnic minorities fit well with the proposition of RDS, which exploits established social networks.

In order to examine methodological utilities of RDS for minority research, we conducted the Health and Life Study of Koreans (HLSK), a Web-based RDS study targeting foreign-born Korean American adults in Michigan and Los Angeles. Although limited, there are probability sample data for this group. Specifically, the American Community Survey and the California Health Interview Survey were used as sources of population benchmarks. We compared the geographic distribution as well as estimates on a series of socio-demographic, health status and care access characteristics between HLSK and benchmarks.

Overall, RDS were promising in geographic coverage. Compare to the benchmarks, HLSK were similar on family type, household size, employment type, and health insurance coverage and type (except for government programs) but over-represented younger, more recent immigrants (hence, lower English proficiency and US citizenship) with higher education and disability, consistently between Michigan and Los Angeles. Particularly, high proportions of recent immigrants in HLSK may imply the effectiveness of RDS for recruiting harder-to-recruit minority subgroups. Despite the encouraging promises, there is much work to do to make RDS a reliable sampling method. Existing RDS-specific estimation methods are ineffective when applied to the real-world data. Recruitment noncooperation proved to intensify operational challenges, a gap in the current literature to make RDS a truly practical methodology.

**Data Collection Mode Change: Going from Face to Face to CATI: The Case of Greece**
*Carsten Broich, Sample Solutions BV*
*Katerina Nikolova, Sample Solutions BV*

**Abstract:** Greece has traditionally been a country utilizing face-to-face data collection for social research projects - examples of that being the PEW Global Attitude Study and the Gallup World Poll.
Provided that 99% of households in Greece have telephone access, and that Greece the highest daily usage of landline phones in Europe (63% from those having access to a landline phone, 90% having access to at least one mobile phone), CATI is a promising option to achieve high coverage and faster turnaround times on completed research [1] . Additionally, many face-to-face studies have neglected the population of Greek islands and focused on the mainland population, resulting in higher coverage error. One of CATI's advantages is reaching the hard-to-reach target audiences, which could add greater insight to a survey, especially because the population living on islands might be significantly different from the mainland population.

This research will explore mode differences and how they contribute to data quality across a variety of metrics including attitudinal and demographic differences, weighting efficiency, and response dispositions. A key research question is to gage the stability of trend data from in-person to phone. Additional analysis will include time of administration, cost and coverage of the island population, compared with the mainland.

[1] Statistics used are from Special Eurobarometer 462: E-Communications and Digital Single Market, fieldwork conducted in April 2017, published in July 2018

## Session: Paradata and Quality

**How Do Timestamps Improve Survey Implementation? Demonstration of PMA Analytics**
*Shulin Jiang, PMA2020, Johns Hopkins Bloomberg School of Public Health*

**Abstract:** Paradata is widely used to monitor interview quality and manage data collection as automatically generated item-level paradata can indicate potential problems with survey items. Studies using large-scale, multi-country sources of paradata from interviewer surveys are limited, however, even though keystroke paradata provides extra details with little additional cost.

PMA2020 is a survey platform that uses smartphone assisted personal interviews (SAPI) to monitor key health and development indicators. It collects a nationally representative sample of data from households, females and health facilities annually in 11 countries. Surveys are carried out by female data collectors - resident enumerators (REs), who are local women from or near the respective enumeration areas. PMA2020 uses Open Data Kit Collect (ODK), a software that facilitates data collection via a mobile-assisted data platform. PMA Analytics runs automatically with ODK when REs conduct interviews on smartphones. The interview process and user interactions are time-stamped and recorded in a log. PMA Analytics data is keystroke data generated from survey submissions and their associated logs.
The advantage of PMA Analytics is its large-scale item-level keystroke paradata. The item-level paradata provides rich details to reflect field activities. The multi-country PMA2020 surveys make it feasible to cross-compare aggregated paradata. Data for this presentation was collected from three out of 11 PMA countries: Ethiopia, Uganda, and India.

In this presentation, we will present four user cases: The first case is how timestamp data was used for the training purpose. The second case is an investigation on an RE's problematic performance on certain questions. The third case presents the difference of interview time among different questionnaire structures. The fourth case shows how Analytics data fits into a broader picture to reflect data quality in an integrative way. All cases together demonstrate how paradata can improve data quality and questionnaire design.

**Comparing and Validating New Methods to Control for Response Biases in Self-Report Educational Data**

*Marek Muszyński, Institute of Sociology, Jagiellonian University*

**Abstract:** Response bias is a general term for a wide array of processes that result in inaccurate or false responses in self-report data (Furnham, 1986). Response biases are perceived as a threat to validity as they lead to lower data quality (Maniaci & Rogge, 2014). Many methods to control for response biases were developed, but none acquired a „golden standard" status. Most of the methods is of unproven utility and have to be implemented before the data is collected.

This calls for a development of methods with a possibility to use also after data collection. Moreover, more validation studies are needed, preferably with use of nonself-report criteria (e.g. cognitive tests). The present research was performed using PISA 2012 dataset, concentrating on the "math familiarity" scale. Numerous methods of identifying biased responses were used (overclaiming questionnaire, long-string analysis, psychometric synonyms, intraindividual response variability (IRV), Cattell's sabotage index, modified fixed individualised chance score (MFIC), Mahalanobis distance, dr* outlier measure, polytomous person-fit statistics).

As the response biases are known to suppress the inter-variable relations, the magnitude of the regression coefficient between the self-assessment math familiarity scale and cognitive math test and the magnitude of the R2 statistic from a multilevel regression were used as a validation of the response bias control methods used.

The IRV and long-string proved to be effective methods of eliminating straightliners, but outlier detecting method yielded hard-to-interpret data as eliminating flagged outliers did not have a noticeable effect on the validation criteria.

The conducted analyses brought new evidence on which response bias methods should be use and to what end, but also yielded new questions, e.g. what cut-offs should be used to identify outliers or how the generated indices should be then put to use in a quantitative analysis of survey results.

**Why Do We Prefer Second Best Quality Indicators?**

*Peter Mohler, Mannheim University*

**Abstract:** "Response bias" is better than "response rate" as a key indicator for sampling quality reporting, at least from a strict methodological point of view. If so, why do we still use response rate as the "gold standard" in quality reports? Similar observations can be made for other parts of a survey such as items or questionnaire design. Two systemic factors can help understanding our preference for second best indicators: 1. Legitimation by procedure (Legitimation durch Verfahren), a term coined by N. Luhmann, that points to the importance of standardized procedures in court procedures and all other societal decision processes. In short, decisions lacking verifyed and standardizes procedurs, will not be

accepted (legitimate). In the case of response bias, we still lack such procedures and hence, this indicator has not and will not be acceptable to users, clients and the research community at large - until we agree on a standard procedure. 2. Tendency not to tackle tough issues, a known psychological phenomenon that groups avoid to tackle hard to get solutions in favor of tackling less relevant side issues that can be handled easily. Tackling response bias would call for hard work on finding the relevant reference values (bias in deviations from well defined benchmark indicators). That would be time and cost intensive, both notoriously lacking in survey research. In the case of response-nonresponse this can be demonstrated by the ever growing number of sub-items reported in survey documentations under the heading "response". The sub-items can be easily counted and classified, but their relevance for sampling quality has not been shown yet.

The paper will illustrate these two factors using the ESS as an example, concentrating at the response-nonresponse issue and a side look at item quality.

As always, simple solutions seldom are, but not so simple solutions will always be.

## Session: Results of the AAPOR/WAPOR Task Force (TF) on Quality of Comparative Surveys

### Overall Goals of 3MC Research

*Timothy P. Johnson, University of Illinois at Chicago*

**Abstract:** In general, 3MC surveys are intended to produce quantitative data that can be compared cross-nationally or cross-culturally. During the past several decades, an impressive array of study and instrument, design, data collection, and data analysis procedures have been developed to support this over-arching goal. While this overall goal is largely shared across the international survey research community, there is less consensus as to our technical ability to achieve absolute equivalence versus close, albeit imperfect, comparability of methods and measures. This presentation will introduce and explore this general controversy, setting the stage for the remainder of the Task Force report.

### Using the Total Survey Error Approach to Assess and Reduce Comparison Error in Cross-National and Cross-Cultural Surveys

*Tom W. Smith, NORC at the University of Chicago*

**Abstract:** Comparability and quality in comparative surveys can be maximized and comparison error minimized by combining the traditional functional equivalence (FE) approach with the total survey error (TSE) approach which has not typically been applied to comparative surveys. TSE has the advantage of being comprehensive, systematic, and rigorous. It covers all components of error, structures those components in an ordered and interrelated manner, and encourages the quantification of error components, when possible, without neglecting qualitative evaluations when quantification is not possible. The TSE paradigm is a valuable approach for comparative studies for several reasons:

- It is a blueprint for designing studies. Each component of error can be considered with the object of minimizing comparison error.
- It is a guide for evaluating error after surveys have been conducted. One can go through each component and assess the level and comparability of the error structures.

- It can set a methodological research agenda for study error and for the design of experiments and other studies to fulfill that agenda.
- It goes beyond examining the separate components of error and provides a framework for combining individual error components into their overall sum.
- By considering error as an interaction across surveys, it establishes the basis for a statistical model for the handling of error across surveys.

For each survey/country in a 3MC survey, each survey contains all error components and if equivalence is not achieved on each component then a comparison error is occurring, e.g., if there is a translation error from the source language used in country A to the target language(s) employed in country B, then there a comparison error in the Wording component. Moreover, besides applying comparison error component-by-component, it can conceptually be seen as applying across all errors to constitute what might be called total comparison error.

**Questionnaire Development in 3MC Surveys**
*Alisu Schoua-Glusberg, Research Support Services*

**Abstract:** The questionnaire development process for 3MC surveys is much more complex than in mono surveys that study just one population. Basic best practice for general survey questionnaire design must be followed, but questionnaire design for 3MC surveys must also take into account potential contributions to validity and measurement error related to ways in which members of different national and cultural groups may differ systematically in how questions are understood and answered. Further, additional operations are required including the adaptation and translation of questionnaire items and other materials developed to convey or request information from survey respondents. A pervasive lack of awareness of the importance of a high-quality translation and of current translation best practices remains a pressing challenge. Multiple forms of pretesting, including cognitive interviewing, are particularly essential for 3MC surveys to test translated and adapted questionnaires. Yet, limited resources means that minimal pretesting is carried out in many 3MC surveys. We discuss these and other key challenges as well as current best practices for 3MC surveys in the areas of questionnaire design, translation, and pretesting.

We next highlight recent methodological and technological advances affecting 3MC questionnaire design as well as the practice of translation and pretesting. For example, 3MC studies are increasingly adopting the use of translatability assessments and advance translations, electronic translation management systems, and in the area of pretesting, cross-cultural cognitive interviewing (CCCI) and online probing. We conclude with discussion of important areas of ongoing research, future directions, and recommendations in this critical stage of the 3MC survey lifecycle.

**Error Sources and Quality in 3MC Sampling and Field Implementation**
*Julie de Jong, University of Michigan*
*Kristen Cibelli Hibben, University of Michigan*

**Abstract:** Countries vary widely in available resources related to sampling design and field implementation; the latter being one of the most challenging components of survey research, particularly in the case of face-to-face surveys conducted across geographically vast and variable terrain. Here, we first highlight error sources which can arise at these two stages and that are unique to or may be more prominent in 3MC surveys, particularly because of the associated impact on comparability. In sampling design, the optimal design for one country may not be optimal for another, leading to

comparison error. In the implementation phase, error arises from the field operationalization of such design decisions alongside other implementation-related sources of error resulting from the non-trivial amount of control over the study to a team of fieldwork supervisors and enumerators which 3MC studies necessarily yield.

We then investigate how the lack of quality control or oversight in a 3MC survey creates a climate where both inadvertent and intentional deviations from standardized processes can have a significant impact on survey quality. Recent technological innovations have significantly changed the menu of methods available for overseeing the process of fieldwork, and include such tools as real-time quality audits, image and audio capturing, and paradata, but implementing these advances has its own challenges, which we will discuss further. We conclude with discussion of potential approaches to minimizing these important sources of error and quality management in 3MC surveys, including data dissemination methods and support of the adoption of the above-discussed practices, as well as ways for the community to continue to share innovations in this rapidly evolving area of survey research. We also discuss recommendations on other best practices in field implementation, in particular when the full use of e-devices for quality control is not an option.

**Survey Documentation in 3MC Surveys**

*Irina Tomescu-Dubrow, Institute of Philosophy and Sociology, Polish Academy of Sciences*

**Abstract:** Survey documentation' refers both to a process and its outcome. As a process, documentation chronicles the lifecycle of a survey, from design, to implementation, data processing, and dissemination. As the outcome of this process, the documentation is, ideally, a publicly-available source of information that generally encompasses a set of documents, which includes the description of the study's methodology, the survey instrument, and the description of the variables. Accuracy of, and access to, *what* gets recorded and *how*, is essential for the informed use of the survey data for rigorous scientific inquiry. Accurate and accessible documentation facilitates transparency and thus allows stakeholders (e.g. the data's producers and consumers) to build on accumulated knowledge. In 3MC surveys, neither the importance of documentation nor its challenges can be overstated. 3MC survey data are collected from different populations (across countries, or ethnic or cultural groups within country), and frequently, over time via repeated cross-sectional designs and, more rarely, panel designs. For understanding both individual country and pooled data, documentation is paramount. At the same time, the complexity of 3MC surveys and the fact that research infrastructures among data producers are unequal leads to the sometimes wildly different documentation standards in content, nomenclature, and access. Available technology to create machine-actionable documentation has yet to help. Few 3MC surveys employ this technology, possibly because it calls for specialized personnel, itself a likely source of additional resource inequality among participating countries in cross-national projects. Altogether, these problems pose substantial challenges to documentation quality, and thus, to data interpretability. To address these challenges, we both systematically discuss them and consider how methodological variables could be developed for inclusion as standard documentation metadata within a 3MC survey dataset.

## Wednesday, March 20th

### Session: Questionnaire Development and Testing

**Testing Branching Techniques and Rating Scales across Countries**
*Aneta Guenova, US Department of State*

**Abstract:** A goal of questionnaire design is to construct questions that are readily understandable to the target population. This intent can be challenged in cross-national research for a variety of reasons, including that response scales may be interpreted differently by different populations. This presentation focuses on branching techniques in the presentation of scale questions and the number of categories offered to explore how modifications in questionnaire design impact marginal distributions and variable relationships. The research also includes insights from cognitive interviewing to help ground the statistical analysis with perspectives from members of the targeted populations.

Using data from a European general-population survey with embedded split-ballot experiments, a number of approaches were tested. One design contrasted a four-point rating scale (very favorable to very unfavorable) with an alternative that asked a directional question (favorable versus unfavorable) followed by an intensity measure (strongly versus somewhat). In another experiment, we compared four- and six-point scales by varying the intensity follow-up item: two-intensity options (very versus somewhat) and three (very versus moderate versus slight). To disentangle "neutral" attitudes, we used another branching technique where we compared four-point scales with five-point scales that included a middle (neither… nor) option. Respondents who picked the middle category were then asked a "leaning" question – that is, whether their opinion was closer to favorable, unfavorable or neutral. While there is considerable room to debate these issues, the research aims to contribute to the literature by (1) describing how variations in question form change the overall conclusions that one would draw from these data; (2) highlighting discontinuities between our findings and commonly accepted best practices; and (3) outlining a program for further cross-national research.

**Establishing Cross-cultural Equivalence for the European Working Conditions Survey: Cross-cultural and Multi-mode Cognitive Pretesting on Employment Status and Job Quality**
*Agnes Parent-Thirion, Eurofound*
*Patricia Hadler, GESIS-Leibniz Institute for the Social Sciences*

**Abstract:** The European Working Conditions Survey (EWCS) seeks to assess the working conditions in all European countries. Since the survey was first fielded in 1990, the heterogeneity of the survey population has strongly increased. This is due to both to the larger number of countries in the European Union and the deregulation of employment, especially in the aftermath of the recession. Thus, the accurate assessment of employment status and subsequent measurement of job quality in a harmonized manner have become more complex.

Eurofound assigned GESIS with the task of coordinating and carrying out cross-cultural cognitive pretesting of several work-related measurement instruments. Cognitive interviews and web probing were carried out with respondents in Germany, Poland, and the UK (web probing only).

Cognitive interviews were used to gain insights on the assessment of employment status, especially for groups of atypical workers, such as people in multi-employment or precarious working situations. The

aim was to determine how atypical groups of workers relate to the construct of employment. Web probing was used to contrast views of employed and self-employed respondents with a large sample. This helped quantify differences in comprehension of questions related to working conditions between these groups.

This paper presents method and results for three constructs, which were examined using both pretesting modes: 'work intensity', 'working time quality' and 'social environment'. We reflect the value of pretesting in more than one country, and pretesting in more than one mode. We will finish by present planned revisions to the next wave of the EWCS to better portray working conditions across European countries in a harmonized fashion.

**Evidence of the Response Processes for Extending Question-and-Answer Models in the Multi-National Web Survey Context**

*Jose-Luis Padilla Garcia, University of Granada*
*Dörte Naber, University Of Granada*
*Isabel Benitez, Universidad Loyola Andalucía*

**Abstract:** The model of the response process to survey questions introduced by Tourangeau (1984, 2018), has been and still is the most prominent model in survey research literature, and has become a kind of "compulsory" reference when working on pre-testing methods. However, extensions of the original model as well as new models have been proposed, such as the extended model of Schwarz and Oyserman (2001), or the ImpExp model of Shulruf, Hattie and Dixon (2008) aimed to account for response sets and cross-cultural differences in the response processes. In our presentation, we will first review briefly existing models of the response processes and secondly, analyze and integrate current theories and findings of survey research. In particular, we will use preliminary results of an ongoing research project to extend the knowledge of response processes for web surveys. We resort to qualitative evidence from Web Probing (WP) method along with psychometrics for survey data collected of 1,000 participants (500 in Germany and 500 in Spain) to single and multi-items of the 8th European Social Survey Round within a mixed-method design. In this presentation, our main focus will be on the social and context effects -the "ecology" of the web survey response processes-, which is often missed by the common views of the cognitive response process. Therefore, we will present and discuss on how to take into account social and contextual factors for modelling response processes in the context of web surveys. In addition, we will discuss on how to integrate web probing qualitative evidence with quantitative results from psychometrics within a mixed research framework for improving cross-national survey data quality.

**Measuring Employment and Earnings among Disadvantaged Youth: Findings from Cognitive Interviews with USAID Youth Workforce Development Program Beneficiaries in Five Countries.**

*Mousumi Sarkar, Well World Solutions, LLC*

**Abstract:** USAID's Office of Education is developing a survey to be used across its programs worldwide to measure employment related outcomes, specifically employment status and earnings pre- and post-intervention, among youth participating in USAID's youth workforce development (WFD) programs. These programs aim to prepare at-risk youths for or place youth in wage employment or self-employment. Activities may include skills and entrepreneurship training, career counseling, or job matching.

Collecting accurate employment and earnings data, particularly among youth, is challenging. Accurate recall of information is one such issue. Our findings indicate, other challenges must be addressed. We find that disadvantaged youth often have difficulty calculating hours worked or amount earned. Additionally, many terms that standard employment surveys, such as the World Bank's Living Standards Measurement Survey, take for granted are completely misunderstood by many of these youth.
We will present our findings from cognitive interviews conducted in Fall 2018 in the Philippines, Kenya and El Salvador, and in January 2019 in Kyrgyzstan and in a low-literacy context (probably Rwanda) with more than 100 youth. We are also piloting the survey in Kyrgyzstan and possibly Rwanda. We will share terms that youth misunderstood, challenges experienced in reporting hours and wages, and how we addressed these issues in revising the survey. We will also discuss our findings regarding the reliability and validity of these items.

## Session: 3MC Case Studies

**Implementing the TSE Framework in a Multinational Context: Lessons from Four National Surveys in the Caribbean**

*Ramiro Flores Cruz, Sistemas Integrales*
*Juan Muñoz - Sistemas Integrales*

**Abstract:** We describe our recent experience with four national household surveys in the Caribbean Region: The Living Standard Surveys of Barbados (2016) and Suriname (2017), The Guyana Quarterly Labor Force Survey (2017) and the Suriname Survey on Violence Against Women (2018).
We faced the challenge of assuring cross-country comparability and targeting different sub-populations for special-interest stakeholders in all countries (farmers in Barbados, labor force participants in Guyana, women 15 to 64 years old in Suriname) while working under unequal levels of local collaboration and with very different sampling frames.

We were nevertheless able to maximize coverage, minimize nonresponse and measurement error, and deliver reliable data only weeks after the end of each data collection period. Some of the applied quality control tools, although relatively unknown in the region, have been used for some time elsewhere, including the integration of computer checks to fieldwork (both with paper questionnaires and CAPI), double-blind back-checks and the monitoring of paradata and survey-specific quality indicators. Less conventional actions, such as auditing of randomized audio recordings, the use of GIS technologies and the automatic encoding of activities and occupations during the interview may become valuable additions to the survey practitioner's panoply.

Moreover, the Barbados and Suriname surveys included a core module on consumption and expenditure to be used for poverty assessment and CPI basket computation. Both surveys used an innovative interviewing approach for the reporting of food consumption that tends to reduce measurement error and may be considered an interesting alternative to the classic diary and recall methods.

**Lessons Learned From Using Paradata to Assess Data Quality and Inform Operational Strategies: Comparative Analyses on National Aging Studies in Thailand and Malaysia**

*Yu-chieh (Jay) Lin, University of Michigan*

**Abstract:** This presentation uses data from two sources: (1) the Malaysia Ageing and Retirement Survey (MARS) in collaboration with the Social Wellbeing Research Centre at University of Malaya and the Survey Research Center at University of Michigan; and (2) the Evolution of Health, Aging, and Retirement in Thailand (HART) in collaboration with the National Institute of Development Administration and the Survey Research Center at University of Michigan. The two studies both adapt same complex instrument designs from the US Health Retirement Study but encounter different challenges such as multi- vs. single languages, centralized vs. decentralized management structures, self vs. proxy interviews, etc. In addition, both studies have more than one eligible respondents selected within each sample unit and interviewers are allowed to interview respondents within the same sample unit without any particular order and to switch among varied interviewing components in a flexible fashion. Paradata is heavily relied upon to monitor interviewers' behaviors.

This presentation focuses on the analysis of keystroke data to assess data quality. We first examine a series of key characteristics between two studies such as sample design, team structure, interviewer and respondent characteristics, etc. These characteristics are then inspected for predictive power against data quality indicators such as interview length, non-response, response changes, etc. Subsequently, in MARS we call back to households that have data quality concerns to verify interviewer's behavior or some survey data collected, among all other information available. Finally, we will present how these analyses of paradata and verification results can be practically applied to improve data quality of interviewer administered surveys

## Session: Harmonizing Panel Survey Data for Multi-Cultural Research

**Comparing Life-Course Data from Multiple National Panel Surveys: Challenges of Ex-post Harmonization**

*Anna Kiersztyn, University of Warsaw*

**Abstract:** This presentation will discuss the specific methodological challenges involved in the ex-post harmonization of panel data on educational and occupational careers from single-country surveys. Such a harmonization effort is a part of a Cross National Biographies – Young (CNB-Young) research proposal which plans to harmonize quantitative longitudinal data on individuals age up to 35 from four countries in order to test hypotheses about the employment trajectories of young adults, with a focus on labor market precarity and its effect on the social structure in a cross-national comparative perspective. The surveys which are to be included in this proposal are: the Polish Panel Survey (POLPAN), the German Socio-economic Panel (SOEP), the British Household Longitudinal Survey – Understanding Society (UKHLS), and the U.S. National Longitudinal Survey of Youth (NLSY79) Young Adults Study. CNB-Young is the first attempt to harmonize detailed longitudinal information on the characteristics of the respondents' successive employment spells (starting from their first job), their full educational histories, income and household composition, and health/well-being indicators. As such, it moves beyond the existing large-scale ex-post harmonization efforts, which either concern cross-sectional data or contain only a very limited set of harmonized panel variables with respect to employment or educational history. This offers potential for substantial methodological advances related to ex-post survey data harmonization, but at the same time poses numerous challenges which will need to be addressed. I will

start with a brief presentation of the background and rationale of the CNB-Young proposal, and then move on to a discussion of the problems involved in the selection of the source variables from each survey, conceptualization of the target indicators, dealing with differences in survey methodology and the country-level institutional and regulatory context.

**Synchronize to Harmonize: Ex-ante Strategy in a Comparative Pseudo Panel Study on Birth Cohorts**
*Ireneusz Sadowski, Institute of Political Studies of the Polish Academy of Sciences*

**Abstract:** The issue of international comparability of survey data applies not only to equivalence of variables but to structural "compatibility" of samples and populations as well. While statistical control of variables regarding social structure might be considered the simplest solution, it is not a perfect one, and more direct comparability can be attained by studying selected cohorts. In my project "Three generations of transformation" I seek to re-utilize data on selected birth cohorts by conducting a survey on Poles born in 1988/1989 and making it ex-ante harmonized with existing Polish data on two earlier cohorts and German data on all three cohorts, thus creating a pseudo-panel. This, mainly replicative, study adds one important assumption to the standard ex-ante harmonization design. For the comparative cohort analysis to be valid, the timing of surveys has to be synchronized on the same age (30-31 in this particular case). As a consequence the design allows to tackle the so called 'APC conondrum' in an internationally comparative framework.

**Constructing Occupational Careers using Panel Survey Data: Harmonization of Multiple Concurrent Jobs through Time**
*Zbigniew Sawiński, Institute of Political Studies of the Polish Academy of Sciences*

**Abstract:** Many surveys record respondent's occupation at a given time through people's main job, usually defined as the one most time-consuming or providing the highest income, rather than via the range of jobs held concurrently. In panel studies, which contain repeated information about jobs at the seams between consecutive survey waves, the decision to ignore occupational data outside main jobs is hard to defend, especially in light of inter-wave panel harmonization. Occupation codes, like any measure, contain errors. Hence, when different panel waves display different codes for the same respondent, three scenarios appear: (1) It is the same job, erroneously coded at least in one wave; (2) it is the same job, but it combines two occupational roles, coded differently (e.g. a teacher who is also a deputy headmaster); (3) these are different jobs, held either simultaneously (and respondent mentioned only one at the time of each survey), or in sequence (i.e. 'actual' mobility). If all these occupation codes are preserved, they can be considered when using panel data to depict occupational careers. I will illustrate various possibilities using the Polish Panel Survey POLPAN, which started in 1988 and is conducted every five years since. We record respondents' occupational data, generally for the 5-years spanning POLPAN waves, and covering all jobs during this time, regardless if performed simultaneously or sequentially. POLPAN shows that for occupational careers divided into one-year intervals, 20 percent of these intervals contain data on more than one occupation. I will employ POLPAN to discuss various solutions for harmonizing occupational data to construct occupational trajectories.