AAPOR/WAPOR Task Force Report on Quality in Comparative Surveys

April, 2021

Chairing Committee:

Lars Lyberg, Demoskop, AAPOR Task Force Chair
Beth-Ellen Pennell, University of Michigan, WAPOR Task Force Chair
Kristen Cibelli Hibben, University of Michigan, Co-Chair
Julie de Jong, University of Michigan, Co-Chair

Contributors:

Dorothée Behr, GESIS – Leibniz Institute for the Social Sciences
Jamie Burnett, Kantar Public
Rory Fitzgerald, City, University of London
Peter Granda, University of Michigan
Linda Luz Guerrero, Social Weather Stations
Hayk Gyuzalyan, Conflict Management Consulting
Tim Johnson, University of Illinois, Chicago
Jibum Kim, Sungkyunkwan University, South Korea
Zeina Mneimneh, University of Michigan
Patrick Moynihan, Pew Research Center
Michael Robbins, Princeton University
Alisú Schoua-Glusberg, Research Support Services
Mandy Sha, www.mandysha.com
Tom W. Smith, NORC University of Chicago
Ineke Stoop, The Netherlands
Irina Tomescu-Dubrow, Institute of Philosophy and Sociology, Polish Academy of Sciences (PAN) and CONSIRT at Ohio State University and PAN
Diana Zavala-Rojas, Universitat Pompeu Fabra, Barcelona
Elizabeth J. Zechmeister, Vanderbilt University, LAPOP

This report was commissioned by the AAPOR and WAPOR Executive Councils as a service to the profession. The report was reviewed and accepted by AAPOR and WAPOR Executive Councils. The opinions expressed in this report are those of the authors and do not necessarily reflect the views of either council.

This report is dedicated to the memory of Lars Lyberg, who has had a profound and lasting influence on our field. He was a generous collaborator, colleague, and mentor, and a great friend.

Table of Contents

Abbreviations used in the report

3MC	Multicultural, Multinational, and Multiregional
AAPOR	American Association for Public Opinion Research
ADQ	Asking different questions
ASQ	Ask-the-same question
ASQT	Ask-the-same questions and translating
AT	Advance translation
CAMCES	Computer-Assisted Measurement and Coding of Educational Qualifications in Surveys
CAPI	Computer-assisted Personal Interviews/interviewing
CCSG	Cross-cultural Survey Guidelines
CESSDA	Consortium of European Social Science Data Archives
CIRF	Cognitive Interviewing Reporting Framework
CRONOS	Cross-National Online Survey
CSDI	Comparative Survey Design and Implementation
CSES	Comparative Study of Electoral Systems
DDI	Data Documentation Initiative
DHS	Demographic and Health Surveys
EES	European Election Studies
EFTA	European Free Trade Association
EQLS	European Quality of Life Survey
ESRA	European Survey Research Association
ESS	European Social Survey
EU-LFS	EU-Labour Force Survey
EU-SILC	EU-Statistics on Income and Living Conditions
EVS	European Values Surveys
EWCS	European Working Conditions Survey
fMoW	Functional Map of the World
GAP	Pew Global Attitudes Project
GDPR	General Data Protection Regulation
GEM	Global Entrepreneurship Monitor
GGS	Generations and Gender Survey
HETUS	Harmonised European Time Use Surveys
IALS	International Adult Literacy Survey
ICT	Information and Communications Technology
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISSP	International Social Survey Programme
LAPOP	Latin American Public Opinion Project
LSMS	World Bank Living Standards Measurement Survey

MCA	Multiple correspondence analysis
NCES	National Center for Education Statistics
ODNI	Office for the Director for National Intelligence
OECD	Organization for Economic and Co-operation and Development
OMB	Office of Management and Budget
PAPI	Paper and Pencil Interviews/interviewing
PCA	Principal Component Analysis
PIAAC	Programme for the International Assessment of Adult Competencies
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPeS	Probability Proportional to Estimated Size
QAS	Quality Appraisal System
QDDT	Questionnaire Design and Documentation Tool
QUAID	Question Understanding Aid
SERISS	Synergies for Europe's Research Infrastructure in the Social Sciences
SHARE	Survey of Health, Aging and Retirement in Europe
SQP	Survey Quality Predictor
SRC	Survey Research Center
SSHOC	Social Sciences and Humanities Open Cloud
SSS	Street Section Sampling
SUSTAIN	Sustainable Tailored Integrated Care for Older People in Europe
TA	Translatability Assessment
TIMSS	Trends in International Mathematics and Science Study
TMT	Translation Management Tool
TRAPD	Translation, Review, Adjudication, Pretest, and Documentation
TRR	True Random Routes
TSE	Total Survey Error
TTT	Train-the-trainer
WAPOR	World Association for Public Opinion Research
WMHS	World Mental Health Survey
WVS	World Values Survey

Executive Summary

Background

Comparative surveys are surveys that study more than one population with the purpose of comparing various characteristics of the populations. The purpose of these types of surveys is to facilitate research of social phenomena across populations, and, frequently, over time. Researchers often refer to comparative surveys that take place in multinational, multiregional, and multicultural contexts as "3MC" surveys (Mneimneh et al., forthcoming).[1] To achieve comparability, these surveys need to be carefully designed according to state-of-the-art principles and standards.

There are many 3MC surveys conducted within official statistics, and the academic and private sectors. They have become increasingly important to global and regional decision-making as well as theory-building. At the same time these surveys display considerable variation regarding methodological and administrative resources available, organizational infrastructure, awareness of error sources and error structures, level of standardized implementation across populations, as well as user involvement. These circumstances make 3MC surveys vulnerable from a quality perspective. Quality problems present in single-population surveys are therefore magnified in 3MC surveys. In addition, there are quality problems specific to 3MC surveys such as translation processes.

The wealth of output from such surveys is usually not accompanied by a corresponding interest in informing researchers, decision-makers, and other users about quality shortcomings. This can lead to understated margins of error and estimates that therefore appear more precise than they actually are. There are also cases where researchers are informed about quality shortcomings but opt to ignore those in their research reports. There are of course many possible explanations for this state of affairs. One is that 3MC surveys are very expensive and the formidable planning and implementation leaves relatively little room for a comprehensive treatment of quality issues. Another explanation is that the survey-taking cultures among survey professionals vary considerably across nations as manifested by varying degrees of methodological capacity, risk assessment, and willingness to adhere to specifications that are not normally applied.

The literature on data quality in 3MC surveys is scarce compared to the substantive literature. There are exceptions, though, including the Cross-Cultural Survey Guidelines developed by the University of Michigan and members of the International Workshop on Comparative Survey Design and Implementation (CSDI). AAPOR has created a cross-cultural and multilingual research affinity group and some 3MC surveys have advanced continuing data quality research programs. Members of the CSDI Workshop have produced three monographs that treat advances in the field of 3MC surveys. There are also scattered book chapters and journal articles that discuss 3MC and quality.

[1] The focus of this report is comparative surveys of individuals in households, which is in line with the missions of American Association for Public Opinion Research (AAPOR) and the World Association for Public Opinion Research (WAPOR). We do not discuss other comparative surveys such as establishment surveys, and agricultural surveys.

The task force has drawn upon this literature and the considerable and varied experience of its members. Many insights into challenges to, and possible solutions for strengthening the quality of 3MC data come from cross-national survey methodology. We note, however, that many societies have cultural and linguistic minorities, with considerable diversity among these groups (Harkness et al., 2014). Therefore, the 3MC issues discussed in the report are also highly relevant to single country multicultural and multiregional survey research, where comparability is also important.

With this context in mind, the main purposes of this task force are to identify the most pressing challenges concerning data quality, promote best practices, recommend priorities for future study, and foster dialogue and collaboration on 3MC methodology. The intended audience for this report includes those involved in all aspects of 3MC surveys including data producers, data archivists, data users, funders and other stakeholders, and those who wish to know more about this discipline. The full Task Force charge can be found in Appendix 1.

Task Force Charge

The Task Force was charged with addressing three main areas:

- What's so special about 3MC surveys?

In Section 2 of the report, we trace the history (2.1), provide examples (2.2) and outline the challenges of 3MC surveys (2.3).

- The notion of quality in a 3MC setting.

The overall goal of achieving quality in a 3MC survey is to minimize error components at the population level as well as across populations. The many challenges associated with achieving this goal across the key stages of the survey life cycle are discussed Section 3. Here the report outlines the issues unique to addressing quality in 3MC surveys including such conceptual issues as differences in commonly-used definitions associated with the term comparability and the extent to which it can be achieved. Challenges arise due to the complex nature of decisions across heterogeneous populations at every stage of the survey life cycle and additional operational steps specific to 3MC surveys, e.g., translation, adaptation and harmonization. Achieving the appropriate balance between standardization across culture, regions, and nations and an appropriate level of localization in the midst of countless ways that survey context can vary must also be addressed at every decision point.

- Basic Design and Implementation Recommendations

Section 4 of the report follows the survey lifecycle: 4.1 Organizational Structure; 4.2 Sampling; 4.3 Questionnaire Design; 4.4 Translation and Adaptation; 4.5 Questionnaire Pretesting; 4.6 Field Implementation; and 4.7 Documentation. Each of these subsections provides an introduction, key operational and design challenges, current best practices, recent innovations, and suggested future directions. We summarize these recommendations below. These are then

followed by recommendations for future research, again, following the survey lifecycle. Finally, we end the report by making the case for a new academic discipline.

Implementation Recommendations

The following are recommendations for each of the stages of the 3MC survey lifecycle. We recognize that some of these are aspirational in nature and may not be feasible in every 3MC survey or in specific study sites within a survey project. Nevertheless, these represent current best practices in order to facilitate quality assessment as well as identifying areas for continuous process improvement. We also recognize that many of these best practices are geared toward cross-national surveys, but many can be applied to within-country cross-cultural surveys as well.

Study Design and Organizational Structure

1. The designated central governing body or sponsor of a 3MC survey should have the capacity to design, implement, train (or coordinate training), monitor and address any challenges to survey quality, as well as have in-depth knowledge of each targeted study population or work with local partners who have such in-depth knowledge.

2. The design and protocols of 3MC surveys should be informed by a combination of TSE, fitness for intended use, and survey quality monitoring to manage the complex and difficult tasks of designing and conducting 3MC surveys.

3. Specifications with accompanying rationale, should be developed by a designated central governing body for every stage of the survey lifecycle, including a process to review and/or approve quality assurance methods, quality control (preferably in real-time), an element of continuous quality improvement, and a quality management system that keeps track of these components.

4. Implementation of mixed-mode designs, novel or reinvented forms of sampling (i.e., nonprobability sampling), and/or inclusion of other forms of auxiliary data (e.g., social media data and government records) should be introduced in such a way as to compare the new method(s) with existing method(s) to investigate differences in quality which would ultimately impact comparability.

5. Organizations conducting 3MC surveys should ensure all local ethics reviews have been completed, approved and are up-to-date.

Sampling

1. Comparable target and survey populations should be defined and documented for each participating 3MC population (often country).

2. Sampling frames in each participating country should be identified and evaluated with consideration given to the quality of available frames.

3. In the absence of an existing sampling frame meeting accuracy criteria, a sampling frame best covering the target population, given budget constraints, should be developed.

4. If the sampling frame is at the level of a household, then a procedure to randomly select respondents from the household should be determined.

5. The sample size necessary to meet the desired level of precision should be determined for each participating country.

Questionnaire design, translation and adaptation, and pretesting

1. Research question(s) or objective(s) should be clearly defined. Survey questions should be drafted after clearly defining concepts of interest to be measured.

2. Subject-area experts, area/cultural specialists, linguistic experts, platform/translation tools technologists, and survey research experts should be a part of the questionnaire development team or process.

3. Some form of translatability assessment, advance translation or a combination thereof should be carried out to make the source questionnaire as easy as possible to translate into other languages and to implement in other cultures.

4. The source questionnaire should be annotated with relevant information for the translation task, e.g., with the intended meaning of key terms and other information deemed crucial for measurement.

5. An analysis plan should be produced relating each survey question to one or more of the research questions.

6. A team translation approach, for instance a TRAPD implementation, should be followed to translate the source questionnaire into target languages. Documentation of particular and/or difficult decisions should be an integral part of the process.

7. An appropriate set of pretesting and/or post-hoc evaluation methods should be used to assess the quality of questions.

8. A documentation scheme should be developed for questionnaire design decisions and changes to the source questionnaire across time for repeat surveys.

9. A documentation scheme should be developed for changes to trend translations (due to errors) across time for repeat surveys.

10. In face-to-face surveys, show cards should be produced for survey items as needed, for use by interviewers in all participating countries following a standard protocol.

Fieldwork (Implementation, monitoring, contact procedures, nonresponse, and paradata)

1. A standard instrument should be developed centrally and then thoroughly evaluated before implementation in all participating countries.

2. A checklist of minimum interviewer candidate requirements should be established, and a comprehensive, standardized interviewer training should be developed and implemented.

3. Interviewer remuneration should be based on hourly pay rates in each participating country.

4. The need and use of incentives for participation should be determined and documented in each participating country as a case-level variable.

5. A standard pilot protocol should be developed and implemented in each participating country.

6. Both computer-generated and interviewer-generated paradata that are critical to collect for quality assessment should be identified, and clear analysis procedures should be developed.

7. A data-driven assessment protocol, based on near real-time quality indicators in the data, for the selection and verification of cases should be established and include thorough documentation for both selection rationale and verification outcome.

Documentation, weighting, and data usage

1. Organizations conducting 3MC surveys should document each stage of the survey lifecycle as it unfolds as well as all input and output harmonization processes resulting in a methodological profile for release alongside the public-use data files (see documentation standards outlined in Appendix 6).

2. The following survey weights should be constructed as relevant for specific project purposes and fully documented: Design or base weights to correct for different probabilities of selection; weights to adjust for undercoverage, nonresponse, and to make weighted sample estimates conform to external values, and; supranational or population size weights to adjust for different national population sizes. A guide should be provided to assist end users with the correct use of survey weights.

3. Producers of 3MC survey data should facilitate data use trainings which include instruction not only on the data structure itself but on the use of documentation materials as well as available paradata.

Priority areas for future research

The comparability of data collected in 3MC surveys is essential for: 1) advancing social science research and training; 2) isolating the role of contextual factors in explaining complex human behaviors and attitude formation; 3) establishing "ranking" of the participating sites (e.g., countries) so that local needs are identified, and local interventions are implemented; and 4) setting strategic resource allocation and policy-making. However, there is a critical need for advancement of knowledge and generation of new scientific theories to address the challenges in obtaining comparability as identified by the practitioners and institutions coordinating major 3MC surveys. To that end, we have outlined a research agenda, with priority areas categorized by areas within the survey lifecycle.

Theory

- Develop a shared language and set of vocabulary for conceptualizing issues of comparability / equivalence / invariance.
- Develop a generalizable model or framework for how cultural variations in cognition, social norms, and language may interact with external variables such as characteristics of the interviewer, the interview setting, the sponsoring and implementing organizations, and the language of the interview among others, to affect survey response and error generating processes.
- Develop theory / guidance on how to design and mount experiments in 3MC surveys.

Study design

- Develop educational materials for sponsors of 3MC survey research about the considerable resources needed for all major design and implementation steps in every country.
- Rigorously test and evaluate cost reduction strategies involving mixed-modes, new technology, multiple data sources (including Big Data), combining probability and nonprobability sampling, and making processes leaner while preserving quality.

Sampling

- The central governing bodies, sponsors or other stakeholders of 3MC surveys working in the same region should seek opportunities to collaborate on initiatives to identify, access, and assess registers and other databases as potentially viable sampling frames.
- Empirically examine the performance of existing sampling frames compared to recent innovations for sampling frame development (e.g., True Random Route (TRR) and Street Section Sampling (SSS)).

Questionnaire development, translation and adaptation,

- Establish and improve existing central resources and databases with tested questions and information on what has been found to work and not work in comparative questionnaire design (e.g., problematic terms, linguistic structures, indicators, lessons learned from major studies) so that lessons can be shared and learned.

- Develop translation quality criteria and methods for assessment, particularly of a quantitative nature.
- Investigate the relative effectiveness of both qualitative and quantitative question evaluation methods and combinations thereof.

Fieldwork implementation and monitoring

- Develop new methods to educate and train interviewers in order to incentivize adherence to study protocol.
- Investigate the most effective approaches to detect both unintentional and intentional deviations from fieldwork protocol from all levels of the survey organization.
- Investigate interviewer and context effects across study countries (social desirability bias, the impact of the perceived social/power distance between the interviewer and respondent, and so on), including measurement metrics, differential impact on data quality, and appropriate analytical methods.
- Identify and/or develop a low-cost mobile data collection software with an integrated sample management system and ability to capture complex paradata while ensuring data security.

Survey quality

- Develop a practical approach for continuous survey quality improvement for 3MC surveys (e.g., Adaptation of the System for Managing the Quality of Official Statistics (ASPIRE) (Biemer et al. 2014)).

Interdisciplinary recommendations

Efforts to foster interdisciplinary research and collaboration, including training courses are needed. Coordination across projects and organizations in the development of new tools and approaches could greatly accelerate theoretical and methodological developments in 3MC surveys, leading to better quality data and increased efficiencies. This requires dedicated funding. The SERISS initiative in Europe provides an example of how such funding has accelerated and advanced the science and practice of 3MC survey research.

Breaking down disciplinary barriers also calls for cooperation at both individual and organizational levels. Organizations like AAPOR, WAPOR, and ESRA, and initiatives such as CSDI and the methodology-oriented research committees of the American and International Sociological Associations, the American and International Political Science Associations, and other stakeholders should form a committee or committees to:

(i) develop strategies to compile and disseminate information about existing resources and best practices in 3MC survey research.
(ii) advance the tools, resources and research in priority areas for future research.
(iii) develop an interdisciplinary training curriculum that would prepare a new generation of specialists in 3MC survey research

3MC survey research should be established as a discipline of its own. This last recommendation demands special justification, since it is critical for the advancement of the science of 3MC research. Given that 3MC surveys are currently conducted by organizations with varying research traditions and experiences regarding survey quality in general, and 3MC survey quality in particular, this report might have a limited effect in some disciplines that are not familiar with AAPOR/WAPOR activities. Frankly, the field of 3MC research is very large with limited collaboration across different research traditions. For example, while theoretical advances in comparative research are made in specific disciplines, including cultural psychology, cultural sociology, linguistics, organizational science, survey methodology, and psychometrics, both the integration and cross-fertilization of these advances with the aim of improving survey data comparability have been limited. While 3MC surveys share the common goal of producing comparable data across many cultures and countries, the lack of communication and coordination among 3MC survey networks as well as between these networks and researchers has hindered opportunities for advancement in improvements to data quality.

Much remains to be done to engage 3MC survey networks, increase connections with researchers conducting cross-cultural research in other fields, particularly in new disciplinary fields such as computational linguistics. A funded effort to increase communication and foster interdisciplinary research and collaboration is urgently needed to advance the science and practice of 3MC survey research.

Further, in order to develop the field, we need to make 3MC research a discipline of its own. So, what does such a development entail? According to Groves (2018), a number of criteria must be fulfilled before a field can declare itself a discipline. The following list is one possible set of such criteria.

a. an academic curriculum should be developed;
b. a professional organization should be created;
c. a scientific journal or a named set of publication outlets should be available to the discipline;
d. the discipline should have a common set of shared values and research principles; and
e. there should be deep ongoing work in knowledge domains.

We cannot yet claim that all these criteria have been fulfilled. There are a few informal interest groups with CSDI at the forefront, research papers are presented at many conferences, and research papers are published in journals that normally cover topics from official statistics to ethnology. Deep ongoing work is indeed being done, but there are problems with outreach across this large field and the diffusion of innovations across disciplines and countries is uneven at best.

According to Groves (2019), all fields need people, people that can be replaced over time. For a field to become a discipline it has to be large enough to attract a critical number of students, faculty, and practitioners. The 3MC literature is comprised of a number of monographs and resources that already now serve as teaching material. What is lacking is a systematic training program, including textbooks for undergraduate and graduate levels. Today scattered single courses are taught in universities, but to move to a product that provides an academic certification, an integration of courses is needed. Also, there is a need for jobs within the

discipline area and here, there appears to be no shortage of opportunities. However, there must be a structured process for training new generations for the field to develop further.

Members of the 3MC field should formalize existing informal groups, form a professional group, and develop this discipline focusing on the criteria above. A group of members selected from this Task Force are in the initiation stages of this process.

1. Introduction

This report discusses best practices for realizing and improving quality in comparative survey research, i.e., projects for which instruments and other aspects of the study and their implementation are "deliberately designed for comparative research" between two or more populations cross-nationally, cross-regionally or cross-culturally (Harkness et al., 2010a, p. 3). The purpose of these types of surveys is to facilitate research of social phenomena across populations, and, frequently, over time. Increasingly, researchers are referring to surveys in multinational, multiregional, and multicultural contexts as "3MC" surveys (Mneimneh et al., forthcoming).[2]

To facilitate comparative analyses, the data must be valid and reliable for the given cultural or national context, as well as comparable across these contexts (Przeworksi & Teune, 1966). This is a formidable challenge. As Harkness et al. (2010a) discuss, comparability should drive design as well as the assessment of data quality in 3MC research. Harkness (2008) argues that, indeed, the pursuit of data quality is simultaneously the pursuit of comparability. In their introduction to the recent text *Advances in Comparative Survey Methods*, Johnson et al. (2019a) state that "3MC methods emphasize the importance and address the comparability of survey data across nations, regions, and cultures" (p. 3).

However, the concept of comparability is highly complex. Academic disciplines employ different definitions and a range of terms for similar but not necessarily identical concepts. The *extent to which comparability can truly be achieved and how to assess the extent to which it is* achieved are subjects of debate. This is discussed further in Section 2.3.

The number and variety of 3MC surveys conducted within the realm of official statistics, academia, and the private sector have grown substantially in recent decades, and with it, their relevance for scientific knowledge production. As Johnson et al. (2019a) note, the potential impact of 3MC survey data on decision-making and knowledge is perhaps more significant than ever.

Many factors have contributed to the rise of 3MC surveys. They include, among others: (i) an increased interest in understanding the consequences of within-country cultural and ethnic heterogeneity (e.g., the impact of race and ethnicity, and their intersection with gender and social class, on education, labor market outcomes, or political participation outcomes, to name just a few), and between-country differences and similarities in causes and consequences, of, for example, economic and political inequalities, health and well-being, migration, and consumer behaviors; (ii) growing emphasis on empirically informed public debate and scholarship, particularly in transitional countries and countries facing rapid political and economic change (Smith, 2010); and, closely, related (iii) to social change itself, with democratization and the relaxing of government restrictions contributing to the spread of comparative survey research in many parts of the world. For example, the AmericasBarometer

[2] The focus of this report is comparative surveys of individuals in households, which is in line with the missions of the American Association for Public Opinion Research (AAPOR) and the World Association for Public Opinion Research (WAPOR). We do not discuss other comparative surveys such as establishment surveys, and agricultural surveys.

(Vanderbilt University) has expanded to 34 countries in the Americas, more than any other comparative project in that region. Pew Research Center's Global Attitudes Survey, founded in 2002, typically includes between 20 and 40 countries a year. In recent years, the European Social Survey (ESS) and the Survey of Health, Aging and Retirement in Europe (SHARE) have added countries in Europe but also beyond. For example, the ESS has been collaborating with the South African Social Attitudes Survey regarding methodology and test fielding of some ESS modules. Also, the Living in Australia panel fielded the ESS core questionnaire in 2019.

However, the growth in 3MC research does not come without challenges. Compared to most single-population surveys, 3MC surveys are much more complicated (see Section 2.3) and the problems associated with their planning and implementation are so demanding that they frequently overshadow quality management and quality assessment activities. This is especially the case when an increasingly larger number of countries with different research experiences and unequal research infrastructures join 3MC survey projects. Broadening geographic coverage and ensuring data quality, particularly in light of financial constraints, can constitute competing pressures (Jowell 1998; Pennell et al., 2017). On one hand, researchers strive for data whose scope is as broad as possible, i.e., greater country coverage. On the other hand, methodological problems, and thus the need for quality control, grow as a function of the number of participating countries – an implication of Sir Roger Jowell's rule to confine cross-national research to the smallest number of countries compatible with a study's intellectual needs (Jowell as referenced in Lyberg, Japec, & Tongur, 2019, p. 1067).

Further, comparative surveys are often rooted in different disciplines with different research traditions, including sociology, psychometrics, marketing, and statistics. While this has resulted in rich disciplinary knowledge, interdisciplinary sharing of accumulated methodological expertise and agreement on common standards on 3MC data quality remain weak.

With this context in mind, the main purposes of this task force, which assembles people from different organizations involved in 3MC survey research, is to identify the most pressing challenges concerning data quality, promote best practices, recommend priorities for future study, and foster dialogue and collaboration on 3MC methodology. The intended audience for this report includes those involved in all aspects of 3MC surveys including data producers, data archivists, data users, funders and other stakeholders, and those who wish to know more about this discipline.

The task force has drawn upon relevant literature spanning a variety of disciplines and types of survey research. Many insights into challenges to, and possible solutions for strengthening the quality of 3MC data come from cross-national survey methodology. Despite this, we note that many societies have cultural and linguistic minorities, with considerable diversity among these groups and their relative sizes throughout the world (Harkness et al., 2014). Therefore, the 3MC issues discussed in this report are also highly relevant to single-country multicultural and multiregional survey research, where comparability is imperative.

The development of 3MC survey methodology is reflected in an increasing number of publications, including two monographs released in 2010 (Harkness et al., 2010b) and in 2019 (Johnson et al., 2019b). 3MC survey related articles have been featured in major journals, including *Public Opinion Quarterly*, the *Journal of Official Statistics*, the *Journal of Cross-cultural Psychology*, and *Quality Assurance in Education* (special issues on the Programme for the International Assessment of Adult Competencies (PIAAC) regarding interview and translation issues). Recent edited volumes and monographs have also included chapters related to 3MC surveys including the International Handbook of Survey Methodology (de Leeuw, Hox, & Dillman, 2008), Hard-to-Survey Populations (Tourangeau et al., 2014), Total Survey Error in Practice (Biemer et al., 2017), and the Sage Handbook of Survey Methodology (Wolf et al., 2016a). The Sage Research Methods Foundations (Atkinson et al., forthcoming) features an entry on 3MC surveys as well.

At the same time, the development of 3MC survey methodology has included professional conferences, such as annual meetings of the International Workshop on Comparative Survey Design and Implementation (CSDI) since 2002, and the International Conference on Survey Methods in Multicultural, Multinational, and Multiregional Contexts.[3] In 2016, a cross-cultural and multilingual affinity group was added at AAPOR and in 2017, a 3MC session track was formally added at the AAPOR annual meeting in recognition of the growth and interest in this area. Academic resources include graduate courses,[4] online short courses (Center for Capacity Building in Survey Methods and Statistics, 2018), and online resources such as the Cross-cultural Survey Guidelines (Survey Research Center, 2016).

In the next section we cover additional background on 3MC research, including a brief history, examples of current 3MC surveys and how they vary, and the fundamental challenges of 3MC surveys. In doing so, we focus on cross-national survey projects. While cross-national surveys constitute one part of the 3MC survey family, their defining feature – to study at least two populations in a given year or over time – is at the core of 3MC research. Most of the challenges international survey projects raise, including with respect to data quality, apply directly to all other 3MC surveys. Section 3 discusses quality in 3MC surveys, existing frameworks, and recent internal and external efforts to assess 3MC survey quality. Section 4 outlines the most pressing design and operational challenges, current best practices, recent innovations, and future directions related to major stages or aspects of 3MC surveys including organizational structure, sample design, questionnaire design, translation and adaptation, questionnaire pretesting, field implementation and monitoring, and documentation. This is followed in Section 5 by a presentation of prevailing issues and the changing survey landscape. Finally, our top-level recommendations are discussed in Section 6.

[3] See http://csdiworkshop.org/

[4] For example, courses have been offered at the University of Illinois at Chicago, the GESIS Summer School in Survey Methodology, and the University of Michigan's Summer Institute in Survey Research Techniques. The Graduate School for Social Research at the Polish Academy of Sciences offers a course on Comparative Survey Methods.

2. Background

Section 2 provides a history of 3MC surveys (2.1), examples of 3MC surveys and how they vary in design, funding, execution and oversight (2.2) and the challenges of 3MC surveys (2.3) including differences across disciplines, etic versus emic, complexities inherent in a 3MC survey, standardization versus localization across the survey lifecyle as well as introduces a TSE framework in a 3MC context.

2.1. History of 3MC surveys

Smith (2010) identifies three distinct periods in the development of comparative survey research. The earliest cross-national surveys were motivated by the context of World War II (see also Mohler & Johnson (2010)). For example, the earliest known attempt to conduct survey research beyond one cultural context were the Strategic Bombing Surveys carried out by the United States government during and immediately after WWII to understand the psychological effects of allied bombing on the morale of civilians in Japan and Germany. According to Smith, comparative surveys during this first period were largely ad hoc, one-time, topic-specific cross-national studies. Deliberately designed cross-national surveys were a rarity (Rokkan, 1969). However, some early examples of cross-national comparative research include the *How Nations See Each Other Study* conducted in nine countries in 1948-49 by Buchanan and Cantril (1953). In addition, a landmark study published during this period was Verba and Almond's *The Civic Culture: Political Attitudes and Democracy in Five Nations* (1963). While initially seen as reflecting the state-of-the-art in cross-national research, subsequent criticism and thinking, most notably from Almond and Verba themselves, on the challenge of achieving equivalence or comparability and the possibility of measurement artifacts, helped researchers to recognize the methodological challenges of comparative survey research and set the stage for future developments in cross-national research (Mohler & Johnson, 2010).

The second phase in the development of comparative survey research saw the rise of sustained and collaborative programs of comparative survey research (Smith, 2010). It was during this period in the early 1970s that multinational survey projects began to emerge. For example, the Eurobarometer, an ongoing attitudes and value orientation survey series, began in 1973, with surveys conducted by the European Commission in 1970 and 1971. The European and the World Values Surveys (EVS and WVS) began in 1981 and the International Social Survey Program (ISSP) in 1985.

Starting in 2002, the third developmental period identified by Smith (2010) was marked by the establishment of the ESS and SHARE, which unlike their predecessors at the time, feature centralized funding for the design, direction, and methodological monitoring of its national surveys. During this period, other comparative surveys began to see central coordination, such as the AmericasBarometer, which began in 2004, and the Arab Barometer, which was initiated in 2005.

The idea of having a central coordinating team or governing body has evolved over the past couple of decades. Previously, it was often thought that countries participating in cross-national studies were able to follow instructions or specifications without much guidance, explanation, or

follow-up. The 1994 International Adult Literacy Survey (IALS) was one of the first cases that demonstrated that this approach had been overly optimistic (Kalton, Lyberg, & Rempp, 1998). France, which ended up last in the IALS country ranking table, protested against the lack of quality control and eventually withdrew from the study. France's concerns were backed by a review team (Kalton et al., 1998) and the European Commission decided to develop a standardized procedure for the conduct of future IALS (Carey, 2000; Lyberg et al., 2018). Other cross-national surveys including the ESS, the World Mental Health Survey, SHARE, and PIAAC have all developed strong central teams and have made progress in the provision of highly detailed specifications and accompanying follow-up procedures (see Pennell et al. (2017) for discussion and examples). Site visits and other meetings aimed at providing clarifications of survey materials and training in survey methods for those responsible for local data collection are also common in these surveys.

The evolution of 3MC survey programs has been accompanied by a parallel development of a 3MC-specific survey methodology. For example, Mohler and Johnson (2010, p. 21) identify five methodological landmarks in 3MC research:

1. The use of *indicators* as the basis for comparison;
2. The recognition of *context* as a relevant determinant for comparison;
3. The application of *translation theory* and theories of meaning to the adaptation/translation of survey instruments;
4. The acknowledgement of *probability multipopulation sampling* as the statistical prerequisite for comparison; and
5. Advances in *statistical methods* allowing for complex modeling such as multilevel analysis or identification of measurement invariance.

Although considerable progress has been made in the development of a 3MC methods discipline, as we outline below, much work remains to be done.

2.2 3MC surveys in practice

Today, 3MC surveys display considerable variation along several dimensions. Examples of major 3MC social, health, and assessment surveys are shown in Table 1.

Important ways in which 3MC surveys vary include their sizes, the sources and flow of funds, which often influence the organizational structure and the extent to which aspects of the survey design are specified. As mentioned, 3MC surveys also represent a broad range of subject areas that are frequently associated with different fields and research traditions, as well as varying levels of awareness of error sources and structures. These aspects are briefly discussed below.

Whether a 3MC survey is regional or global is a key determinant of size in terms of the number of participating countries. Participating countries in regional surveys typically number in the 20s to 30s. Examples include the ESS, in which 30 countries participated in Round 9 in 2018, and SHARE, which covered 28 countries in Wave 7 in 2017. Round 4 (2014-2016) of the Asian Barometer Survey network included research teams from 14 East Asian countries and five South Asian countries. The AmericasBarometer expanded to cover 29 countries in the Americas in the

2016/2017 Round, and now covers 34 countries total in its database. Surveys conducted in the European Union (by the National Statistical Institutes) include all EU countries, European Free Trade Association (EFTA) countries and at times, EU candidate countries.

Obviously, the 'global' surveys generally cover considerably more countries. The WVS is carried out in up to 75 countries, the Gallup World Poll has been conducted in more than 160 countries, and, as noted above, Pew Research Center's Global Attitudes Survey typically includes between 20 and 40 countries a year. Surveys carried out by the Organization for Economic Cooperation and Development (OECD) have mostly focused on their member countries, which were historically industrialized countries but have expanded in recent years to also include some countries with emerging economies. Twenty-eight countries participated in the first round of PIAAC at some point between 2008-2013, and 72 countries in PISA in 2015.

Table 1. Examples of Current 3MC Surveys

Social Surveys	
Global	Comparative National Elections Project, Comparative Study of Electoral Systems (CSES), Gallup World Poll, Gallup International Voice of the People, Generations and Gender Survey (GGS), Global Corruption Barometer, Global Entrepreneurship Monitor (GEM), International Social Survey Programme (ISSP), Luxembourg Income Study*, Luxembourg Wealth Study*, Pew Global Attitudes Survey, World Bank Living Standards Measurement Survey (LSMS), World Values Survey (WVS)
Regional	AfroBarometer, AmericasBarometer, ArabBarometer, Asian Barometer, Caucasus Barometer, Central Asian Barometer, East Asian Social Survey (EASS), EuroBarometer, European Crime and Safety Survey, European Election Studies (EES), European Quality of Life Survey (EQLS), European Social Survey (ESS), European Values Survey (EVS), European Working Conditions Survey (EWC), Eurosystem Household Finance and Consumption Network*, EU-Labour Force Survey (EU-LS)*, EU-Statistics on Income and Living Conditions (EU-SILC)*, Harmonised European Time Use Surveys (HETUS)*, LatinoBarometer, South Asian Barometer, Transatlantic Trend Survey
Health Surveys	
Global	Demographic and Health Surveys (DHS), Family and Fertility Survey, Global Adult Health Survey, USAID Act to End Neglected Tropical Diseases, World Health Surveys, World Mental Health Survey
Regional	European Health Interview Survey, Survey of Health, Aging, and Retirement in Europe (SHARE)
Educational Surveys	
Global	Programme for International Assessment of Adult Competencies (PIAAC), Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS)
Regional	Adult Education Survey

*Indicates that the survey is post-harmonized.

3MC surveys are funded in many different ways and large survey efforts often involve multiple funding sources, which can affect quality ambitions. ISSP investigates current social science topics in each participating country. Each survey organization has funded all of its own costs; there are no central funds. SHARE has one central source of funding and a centralized administrative unit, as do EQLS and EWCS. In the ESS, all participating countries – Member and Observer countries of the European Research Infrastructure Consortium – contribute to the central coordination costs by a basic membership fee and an additional amount, calculated according to the GDP of each country. In addition, each country covers the cost of fieldwork and national coordination (European Social Survey, 2020). The World Mental Health Surveys receive support from the U.S. government and a number of foundations, among other sources. In addition, each participating country has had its own sources of funding. EU surveys are mostly funded by the member countries. Pennell et al. (2010) and Cibelli Hibben et al. (2016) provide additional examples and further discussion of how a number of existing 3MC survey programs have been funded.

The source (or sources) and flow of funding frequently dictate the organizational structure of a study. As Pennell et al. (2010) discuss, funding obtained through a central source usually means that the organizational structure is determined by the organization in receipt and control of these funds. Organizational structures for 3MC surveys can be seen as laying on a continuum in terms of the locus of control. The locus of control may be centralized (all design and operational decisions controlled by a central governing body) or decentralized (each country makes its own operational decisions while adhering to the study design protocols set by the centralized team or a governing body) (Cibelli Hibben et al., 2016). At one end of the continuum, for a study that is decentralized, just a source questionnaire or a list of variables may be provided, and the details of implementation are left up to the participating countries and service providers who deliver the requested data. The other extreme can be represented by surveys such as ESS, SHARE, WMHS, TIMSS, PIAAC and the AmericasBarometer, each with highly centralized infrastructures with a continuously improving machinery for survey planning, implementation, and monitoring adherence to specifications (Pennell et al., 2017). A variation on a centralized infrastructure involves commissioning all of the field work to one company which in turn contracts and coordinates the in-country data collections. This is the case with EQLS and the Eurobarometer.

Academic traditions can vary widely in their scientific approach to comparative survey methods and, in some cases, can be quite entrenched. Two major types of variation can be readily observed. One is the remarkable variation that exists in survey methodology know-how and capacity among countries, sometimes even between neighboring countries or suppliers within a country (Lyberg et al., 2019). The other type of variation stems from the features of research traditions associated with specific subject-matter areas. In assessment surveys, for example, a lot of energy goes into developing the psychometric items and less into developing other survey instruments, such as background questionnaires.

Harmonization in 3MC survey research

A defining feature of 3MC surveys is the need for some form of harmonization. Harmonization is a generic term for procedures aimed at achieving or at least improving the comparability of answers that respondents who are surveyed in different populations or periods provide (Granda

& Blaszcyk, 2016). Depending on whether researchers plan a comparative study, or seek comparability of existent data not designed *a priori* as comparative, the literature distinguishes between input harmonization, *ex-ante* output harmonization, and output (i.e., *ex-post*) harmonization (Ehling & Rendtel, et al., 2006, p. 1-2; Granda, Wolf & Hadorn, 2010).

3MC surveys feature primarily a mix of input and *ex-ante* output harmonization. Data producers apply a variety of methods before fieldwork begins (input harmonization), during data processing (*ex-ante* output harmonization), or at both these stages (Wolf et al., 2016b). Input harmonization resembles standardization. For example, in a cross-national study, national teams would agree, from the beginning, on common concepts (i.e., standardization of definitions), common measurement of the concepts (standardization of instruments), common questions based on a common source questionnaire, common training (e.g., national coordinators; translators), and common technical requirements (e.g., minimum response rate) (Ehling, 2003).

By contrast, in *ex-ante* output harmonization national teams would agree on a common target variable and a common measurement pattern, but use country-specific survey items to collect the data (Granda et al., 2010; Hoffmeyer-Zlotnik, 2016; Kallas & Linardis, 2010). Once the data are collected, variables are recoded following the harmonized coding schema. As an example, consider the International Standard Classification of Education (ISCED) harmonized measure of education levels: it is obtained via "mapping" of national classifications of education. The underlying assumption in *ex-ante* output harmonization is that, for concepts that are common for different populations, comparable estimates can be obtained despite the lack of a common ground such as similar essential survey conditions (Baldacci, Japec, & Stoop, 2016; Eurostat, 2019).

Some 3MC research (e.g., ESS, ISSP) combine input harmonization with *ex-ante* output harmonization for selected data elements (Kallas & Linardis, 2010). For example, ESS features substantial input harmonization but also provides several *ex-ante* output harmonized variables using international standard classifications (e.g., ISCED, post-coding of respondents' occupations using the International Standard Classification of Occupations, ISCO (European Social Survey, 2018a).

Ex-post harmonization is mostly the purview of secondary users. However, some data producers also apply harmonization methods *ex-post*, to already released files that are not comparable by design, to integrate them into datasets suitable for comparative analysis. Examples include the Luxembourg Income Study, the Multinational Time Use Study, and the Cross-national Equivalent File.

Ex-ante and input harmonization strategies are preferred and recommended for 3MC surveys, since they allow comparability issues to be addressed, first at the design stage, throughout data collection, and then during data processing, when creating harmonized files. However, even when applied properly (e.g., strong input harmonization and monitoring adherence to specifications) risks to survey data quality persist. Challenges are even greater when *ex-ante* harmonization is weak. We discuss many of these issues throughout the rest of the report.

2.3 The fundamental challenges of 3MC surveys

3MC surveys face a number of fundamental challenges, including such conceptual issues as differences in commonly-used definitions associated with the term comparability, what constitutes comparability, and the extent to which it can be achieved. Additional challenges arise due to the complex nature of decisions across heterogeneous populations at every stage of the survey life cycle and additional operational steps specific to 3MC surveys, including translation and adaptation. The many challenges associated with key stages of the survey life cycle are discussed in more detail in Section 4. Also critical is the issue of achieving the appropriate balance between standardization across culture, regions, and nations and an appropriate level of localization in the midst of countless ways that survey context can vary. The overall goal is to minimize error components at the population level as well as across populations. This is discussed further below.

Conceptual challenges

The roots of 3MC research can be traced to contributions made by numerous disciplines (Figure 1). Over the past several decades, this collective body of literature has provided the foundation for the development of 3MC research principles. Appropriately, this work has placed considerable emphasis on conceptual and methodological strategies for developing and verifying equivalence or comparability of survey measurements across cultures, regions, and nations. Yet, it has been recognized for some time that the terminology employed when addressing these issues is itself neither equivalent nor comparable, perhaps as a consequence of its origin in these multiple disciplines and research traditions. Approximately 20 years ago, a review of this multi-disciplinary literature concluded that literally dozens of forms of equivalence were being discussed in practice (Johnson, 1998). These discussions often employed different terms to denote the same underlying concept, and also used similar terms to reference differing equivalence concepts. Since that time, the variety of conceptualizations of equivalence in this literature has continued to expand, to more than 90 as of today (see Johnson (2019) and Table 2 reproduced below). As a consequence, considerable confusion regarding the meanings of competing and overlapping definitions of equivalence remains. This ambiguity, lack of consensus, and absence of shared terminology is now a serious barrier to continued progress in 3MC research.

Figure 1. Disciplines contributing to 3MC survey research

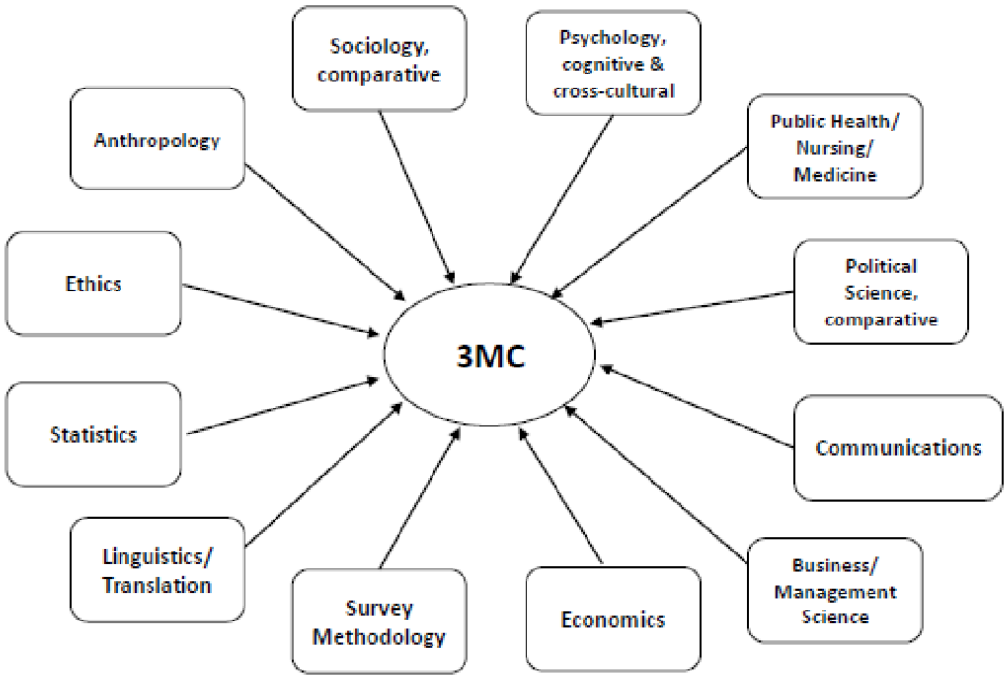


Table 2. Forms of equivalence discussed in the research literature[5]	
Administration equivalence (Buil et al., 2012)	Literal equivalence (Verba, Nie & Kim, 1978)
Approximate equivalence (Davidov et al., 2015)	Meaning equivalence (Prince & Mombour, 1967)
Calibration equivalence (Craig & Douglas, 2000)	Measure equivalence (Craig & Douglas, 2000)
Category equivalence (Buil et al., 2012)	Measurement equivalence (Buil et al., 2012)
Categorical equivalence (Craig & Douglas, 2000)	Measurement instrument equivalence (Zavala-Rojas et al., 2019)
Communicative equivalence (Saule & Aisulu, 2014)	Measurement model equivalence (Hox et al., 2010)
Complete equivalence (Verba, Nie & Kim, 1978)	Measurement unit equivalence (van de Vijver & Leung, 1997)
Conceptual equivalence (Harkness, 2003)	Metaphorical equivalence (Dunnigan et al. 1993)
Configural equivalence (Hox et al., 2015)	Metric equivalence (Craig & Douglas, 2000)
Connotative equivalence (Veselinova, 2014)	Model equivalence (Singh, 1995)
Construct equivalence (van de Vijver & Leung, 1997)	Motivational equivalence (Triandis, 1972)
Construct operationalization equivalence (Hui & Triandis, 1983)	Normative equivalence (Behling & Law, 2000)
Content equivalence (Tsai et al., 2018)	Operational equivalence (Stevelink & van Brakel, 2013)
Context equivalence (Flaherty et al., 1988)	Paradigmatic equivalence (Špirk, 2009)

[5] Table 2 references are included separately in Appendix 2.

Contextual equivalence (Elder, 1973)	Partial equivalence (Hox et al., 2015)
Credible equivalence (Teune. 1990)	Pragmatic equivalence (de Jong et al., 2019)
Criterion equivalence (Flaherty et al., 1988)	Procedural equivalence (Johnson, 1998)
Cross-cultural equivalence (Tsai et al., 2018)	Pseudo equivalence (van Deth, 1998)
Cross-national equivalence (Davidov et al., 2014)	Psychological equivalence (Eskensberger, 1973)
Cultural equivalence (Devins et al., 1997)	Psychometric equivalence (Devins et al., 1997)
Data collection equivalence (Sekaran, 1983)	Referential equivalence (Kenny, 2001)
Data equivalence (Buil et al., 2012)	Relational equivalence (Ellis et al., 1989)
Definitional equivalence (Eyton & Neuwirth, 1984)	Relative equivalence (Frey, 1970)
Denotive equivalence (Veselinova, 2014)	Response equivalence (Buil et al., 2012)
Direct equivalence (Frey, 1970)	Sampling equivalence (van Herk et al., 2005)
Dynamic equivalence (Kashgary, 2011)	Scale equivalence (Anderson et al., 1993)
Exact equivalence (Kashgary, 2011)	Scalar equivalence (van de Vijver & Leung, 1997)
Experiential equivalence (Sechrest et al., 1972)	Scoring equivalence (van Herk, 2000)
Factor equivalence (Dressler et al., 1991)	Semantic equivalence (Tsai et al., 2018)
Factorial equivalence (Hox et al., 2010)	Situational equivalence (Kashgary, 2011)
Formal equivalence (Kashgary, 2011)	Statistical equivalence (Zavala-Rojas et al., 2019)
Full equivalence (Veselinova, 2014)	Stimulus equivalence (Kleiner, Pan & Bouic, 2009)
Full score equivalence (van Herk et al., 2005)	Stylistic equivalence (Veselinova, 2014)
Functional equivalence (van Herk et al., 2005)	Syntactic equivalence (Kohn & Słomczyński, 1990)
Grammatical equivalence (Leonardi, 2000)	Syntagmatic equivalence (Špirk, 2009)
Grammatical-syntactical (Sechrest et al., 1972)	Structural equivalence (van Herk et al., 2005)
Idiomatic equivalence (Sechrest et al., 1972)	Technical equivalence (Herdman et al., 1997)
Indicator equivalence (Kuechler, 1987)	Text equivalence (Saule & Aisulu, 2014)
Institutional equivalence (van Herk et al., 2005)	Text normative equivalence (Veselinova, 2014)
Instrument equivalence (Singh, 1995)	Textual equivalence (Leonardi, 2000)
Instrumentation equivalence (van Herk, 2000)	Theoretical equivalence (Teune, 1977)
Inter-cultural equivalence (Feldkircher, 1998)	Translation equivalence (Craig & Douglas, 2000)
Interpretive equivalence (Johnson, 1998)	Translational equivalence (Hui & Triandis, 1983)
Item equivalence (Borg & Shye, 1996)	True-score equivalence (Riordan & Vandenberg 1994)
Language equivalence (Herdman et al., 1997)	Verbal equivalence (Adams-Esquivel, 1991)
Lexical equivalence (Blumer & Warwick, 1993)	Vignette equivalence (Elder, 1976)
Linguistic equivalence (Iyengar, 1976)	Vocabulary equivalence (Sechrest et al., 1972)

Mohler and Johnson (2010) have also challenged the appropriateness of conceptual reliance on the term "equivalence" in 3MC research more generally. They believe equivalence is a philosophical term that implies it is possible to measure identical dimensions across cultures. They suggest that perfect or absolute equivalence across cultures or nations is more of an aspiration than an achievable goal. Instead, they advocate reliance on the objective of methodological comparability, which they believe to be less absolute and which can be more realistically realized in practice. From this perspective, comparability is conceptualized as the possibility of measuring the similarity, or measurement overlap, of well-defined characteristics of two or more objects under observation using scientific methods. Verba, Nie, and Kim (1978) have also previously suggested that complete equivalence is a hypothetical achievement that may be unattainable in practice.

A final conceptual challenge inherent in 3MC surveys is the extent to which comparability can be truly achieved for at least some concepts. As Smith (2019a, pp. 29-30) has noted:

> Cultural traits are often described as either emic, referring to those that are culture specific or close to being societally unique, versus etic, which describes aspects seen as universal that are "understood in a consistent manner across cultures and national boundaries (i.e., to the extent that they have interpretive equivalence) (Johnson, 1998)." Some concepts are so emic that they are even hard to formulate in other languages for other cultures. For example, "giri" is an indigenous Japanese concept having to do with social interaction, duty, and obligation that at least one researcher, Ruth Benedict (1946), described as follows, "There is no possible English equivalent and of all the strange categories of moral obligation which anthropologists find in the culture of the world, it is one of the most curious." Similarly, the American concept of "hard work" is readily understood in the United States as a chief means by which individuals can advance and improve their lot in life. In other countries the concept is not as clear and pervasive and has been misunderstood to mean "work that is difficult to do" or that people can advance by taking on difficult work, perhaps because there is higher pay for such tasks.

Researchers are often drawn towards the etic rather than the emic, since how can one compare what is unique and does not exist across countries? But that can be a mistake. If one only examines the etic and ignores the emic, one both creates cross-national images of societies that are more homogenous than they actually are and generates a more superficial portrait of each individual society.

One useful approach to bridging the emic/etic cultural divide is to develop items that combine the two. This etic-plus-emic approach is useful when the common core is adequate for direct comparisons. For example, a study of obedience to authority in the United States and Poland had five common items plus three country-specific items in Poland and four in the US (Miller, Słomczyński, & Schoenberg, 1981). This allows both direct cross-national comparisons as well as more valid measurement of the construct within countries, and presumably better measurement of how that construct works in models (Fetvadjiev et al., 2015).[6]

Particularly challenging is when substantive differences interact with methodological or measurement differences. For example, Uskul and Oyserman (2006) and Schwarz, Oyserman, and Peytcheva (2010) show how substantive differences between East Asian collectivist societies and Western individualist societies lead to differences in how information is processed and how survey questions are responded to. Similarly, it has been frequently observed (Smith, 2010) that East Asians in general and the Japanese in particular avoid extreme responses to questions. It has not been determined if the avoidance of extreme responses is tied to translation biases, differential response styles, real cultural differences, or some combination of methodological and

[6] If the core items and the core plus country-specific items formed reliable scales that both showed the same basic relationships in models, then results would be clear and robust. The appearance of different patterns for the core and country-specific items would of course raise questions about cross-national validity.

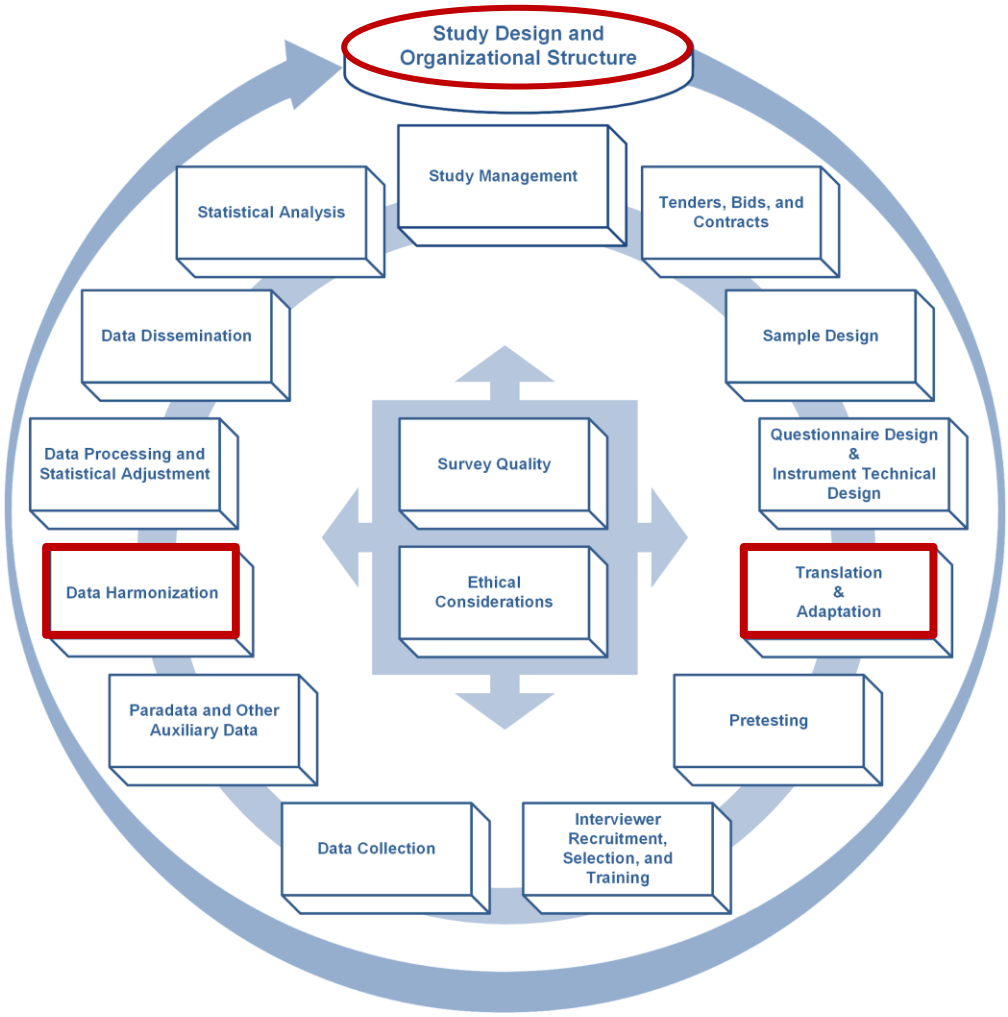
substantive factors. Likewise, Andreenkova (2015) shows that differential acquiescence may explain cross-national differences in attitudes.

There still remains a lack of consensus regarding the meaning of the fundamental language pertaining to the comparability or equivalence of survey measures that are being developed, collected, and analyzed across nations, regions and/or cultures. Developing a shared language is essential to further advancement. Doing so will be challenging, given the diversity of participating disciplines, each of which has been historically "siloed," independently developing their own sets of practices for conducting comparative survey research. A recent paper by Padilla, Benitez, and van de Vijver (2019) provides a good example of work intended to reconcile and unify the terminology and conceptualization of equivalence across disciplines. Achieving consensus regarding the fundamental terminology of 3MC research should be viewed as a long-term goal that will continue to require considerable thought, effort, and cooperation.

Additional operational steps

While the success of 3MC surveys hinges on the comparability of data across many cultures and countries, the practical challenges inherent in implementation, documentation, survey quality assessment procedures, and associated criteria are far more complex than in single-population surveys. Quality problems present in single-population surveys are magnified in 3MC surveys and also new quality problems specific to 3MC surveys are introduced. 3MC surveys face additional challenges and complexity at all stages of the survey lifecycle. Compared to single population surveys, 3MC surveys also require additional operations, such as translation and the adaptation of questions and other survey materials, so that intended meanings are preserved across cultural groups or nations. 3MC surveys, particularly those that are cross-national, have an extra layer of overall study design, organizational structure and harmonization, in addition to aspects that must be considered for any single country survey. These additional operational stages are highlighted in a modified version of the survey lifecycle diagram for 3MC surveys from the Cross-cultural Survey Guidelines (Survey Research Center, 2016) shown in Figure 2.

Figure 2. 3MC survey lifecycle



Standardization and localization

The variation in 3MC survey contexts is considerable (Pennell et al., 2010; Pennell & Cibelli Hibben, 2016). Therefore, collecting comparable data in a 3MC context is a highly complex task, in which one can expect to encounter many challenges. Even in a single-country survey, the target population may not be linguistically, ethnically, or culturally homogenous. Such heterogeneity may manifest itself through a number of dimensions. For example, as Pennell et al. (2010) note, language (e.g., local languages may not have a standard written form; varying respondent literacy rates), geographic topography (e.g., remote islands, deserts, or mountainous regions), weather and seasonal impediments (e.g., winter/summer, monsoons), national and religious holidays (e.g., the Christmas season, Ramadan), or political upheavals may make the harmonization of fielding times across different countries impractical. Moreover, some populations may be inaccessible because of migration patterns or interviewer safety concerns, or they may be only accessible under special circumstances (e.g., miners in camps, or populations

in which part of the population goes on long hunting or fishing trips). See also Pennell and Cibelli Hibben (2016) for a detailed discussion of the numerous aspects of survey context affecting survey design features across cultures and nations.

Countries also vary considerably in survey research infrastructure, experience with various methodologies and technologies, and in their laws, norms, values, and customs pertaining to data collection and data access. Certain modes of administration may not be feasible. In addition, nonresponse levels and biases will likely vary due to differences in cooperation and ability to contact respondents (Stoop et al., 2010). Finally, some countries officially prohibit survey research (e.g., North Korea) or severely restrict data collection on some topics (e.g., see the recent report on *Freedom to Conduct Opinion Polls* (Frankovic, Johnson, & Stavrakantonaki, 2017)).

While a survey conducted in a single country might face one or more of the challenges mentioned above, the probability of encountering several of these is much higher in a large-scale 3MC study. "What is atypical in the one-country context often becomes the norm in 3MC contexts. Moreover, the assumed homogeneity and common ground that may, broadly speaking, hold for a single-country study contrasts with the obvious heterogeneity of populations, languages, and contexts encountered in multinational studies." (Pennell et al., 2010, p. 270).

To maximize comparability, a strict standardization of the study design and implementation protocols across populations is not always possible nor even desired. For 3MC surveys, the challenge is how to achieve the appropriate balance between standardization across cultures, regions, or nations and an appropriate level of localization to minimize error components at the population level and across populations. Unfortunately, as Pennell et al. (2017) note, there is no set model or framework to guide comparative surveys, although some large comparative surveys are moving toward setting minimum requirements, monitoring these throughout the survey lifecycle, and documenting the outcome. In practice, because of the need to balance standardization with localization, absolute input harmonization is not possible or even desirable for some survey processes due to methodological and administrative resources available or because it would not be suitable in one or more of the cultural contexts. Thus, to achieve comparability, a blend of requirements and a certain level of flexibility is needed (Pennell et al., 2017).

There can also be considerable variation in cultural contexts during the operationalization of a survey across populations in different countries, potentially leading to statistical variation that is at least partially attributable to measurement error rather than real differences. Responses can be affected by cultural variation in issues of understanding and interpreting a question, challenges of recall, and variation in judgement formation, presence of others, response mapping, and response editing processes (Johnson, O'Rourke, & Chavez, 1997; Johnson & Braun, 2016). However, ascribing different response patterns to differences in cultural backgrounds alone is not supported by evidence provided by Beullens and Loosveldt (2016) and Loosveldt and Beullens (2017). Their analyses of data from the European Social Survey show that there can be large differences among countries arising from interviewer effects that cannot be totally ascribed to language differences or respondent differences. There are also occasionally fairly large differences among

interviewers within the same country, which again cannot be totally ascribed to language differences or respondent differences (Schnell & Kreuter, 2005).

Cultural norms also play a role in social interactions and communication patterns and can contribute to measurement error (see Johnson and Braun (2016) for a review of this literature). Additionally, the impact on data from social desirability bias occurring within an interviewer-respondent interaction can vary significantly across countries. For example, one recent analysis has shown a strong association between interviewer attitudes and respondent attitudes, and another has demonstrated evidence of variance in interviewer behavior with regards to requesting a private setting in an interview and potential consequences to data quality (de Jong, Mneimneh, & Moaddel, 2017; Mneimneh et al., 2018). Lastly, structural differences across countries can lead to significant measurement error, particularly when collecting sociodemographic data such as education, income, and occupation, whose response categories may vary widely and necessitate significant harmonization before and/or after data collection (Braun & Mohler, 2003; Braun & Müller, 1997; Hoffmeyer-Zlotnik & Wolf, 2003; Schneider, 2007; Schnepf, 2018).

To help conceptualize error in the 3MC context, Smith (2011) and Pennell et al. (2017) have expanded the traditional Total Survey Error (TSE) framework to include the concept of "comparison error". Originally defined by Smith (2011), comparison error is the error introduced across each stage of a 3MC survey as well as the aggregate of error across all stages. Smith's Figure A, shown in Appendix 3, delineates errors in a series of components and sub-components starting with Sampling and Non-Sampling Error and then breaking the error down into 35 error components shown in the right most boxes in each sub-division of errors. From each box come two types of error, namely, variance or random error (shown by the black lines) and bias or systematic error (shown by the red lines). Figure B, also shown in Appendix 3, shows that for each error component in one survey (say in Country 1), there is a matching component in a second survey (say in Country 2). If there were 30 countries in a comparative study, then there would be 30 stacked error boxes for each of the error components. In a comparison of two countries A and B, the step of optimizing a measure requires, for both countries/languages, great care and is sometimes very difficult but is otherwise straightforward. But with countries A, B, and C, the best choice for A and B may not work for C and changing A and B to best work with C may no longer work well for A and B. Now taking countries A to Z, it is clear that the task of optimizing an item could be a very difficult undertaking or not achievable at all.

Smith's Figure A and Figure B demonstrate the numerous potential sources of error and the incredible complexity that very quickly develops in a cross-national context as the number of countries increases. Also, key to consider is the frequent interaction between different TSE components (see Smith (2011; 2019a) for further discussion).

Pennell et al. (2017) discuss key challenges particular to 3MC surveys for each of the Total Survey Error (TSE) representation and measurement error components. The TSE framework in Pennell et al. (2017), shown in Appendix 4, Figures A and B, links error sources to the key stages of the survey process – design, implementation, and evaluation – and identifies, for each error component (e.g., coverage error, sampling error, and measurement error), key potential sources of error that may contribute to TSE in individual populations and may present particular challenges in standardizing design and implementation (or establishing suitable localized

equivalents) across populations, thereby potentially increasing comparison error. It also incorporates the dimensions of cost, burden, professionalism, ethics, and other design decisions that frequently impose constraints on 3MC survey design and have an important influence on 3MC survey quality. Importantly, the authors highlight that the effect of various error sources are statistic-specific and therefore need to be considered not at the level of the survey but, ideally for each measure of interest.

For most of these challenges, there is no consensus among 3MC practitioners and methodologists as to how they should be handled. It is not just a matter of mitigating and controlling the errors. Many trade-off decisions are also necessary.

3. Quality and comparability in 3MC surveys

Section 3 further discusses quality and TSE and introduces the concept of comparison error, as well as fitness for intended use and how these concepts apply to 3MC surveys.

The specific challenges and characteristics of 3MC surveys necessitate a unique approach to addressing survey quality. Indeed, each 3MC survey is shaped by factors such as the number of countries, the capacity of partners and resources available, the locus of control, the funding mechanism, overall budget, and even political decisions. However, there have been several attempts to identify those factors considered most essential in a comparative survey. For example, a methodological report on the International Adult Literacy Survey (IALS) includes a primary conclusion that a *strong infrastructure* is necessary to maintain adherence to requirements ensuring quality (Carey, 2000), and this sentiment is underscored by others as well (Kalton et al., 1998; Lyberg et al., 2019; Murray, Kirsch, & Jenkins, 1998; Pennell et al., 2017).

In addition to a strong infrastructure, there is general agreement on the necessity for robust quality assurance (QA) and quality control (QC) systems. Such protocols would include a comprehensive set of specifications that define the design implementation and the quality assurance steps necessary for all participating countries, a QC system to verify adherence to specifications, both a central team and strong local partners responsible for implementation and maintenance of the QC system, and, lastly, the location and status of the central team within a greater infrastructure that can administer methodological capacity building and provide quality management.

However, the lack of a comprehensive set of universally accepted standards and best practices to guide 3MC surveys has led to concern about their quality. As discussed above, current 3MC surveys display considerable variation regarding methodological and administrative resources available, organizational infrastructure, awareness of error sources and error structures, level of standardized implementation across populations, and user involvement. These circumstances make 3MC surveys vulnerable from a quality perspective. As 3MC surveys expand into new areas, with new funding sources, it becomes even more important to consider issues of quality and to outline best practices that extend beyond the recommendation for a central organizational structure.

Three different approaches to survey quality are most widely applied and/or adapted to the 3MC context: 1) total survey error and comparison error; 2) fitness for intended use; and 3) monitoring survey quality. In a 3MC setting, quality is achieved by reducing total survey error of important estimates across all targeted populations while retaining the ability to compare across these populations.

Total survey error (TSE) and comparison error

Total survey error (TSE) (e.g., Biemer 2010, 2016; Groves et al., 2009; Groves & Lyberg, 2010; Lyberg & Weisberg, 2016) is widely accepted as the organizing framework in the design, implementation, and evaluation of single-country surveys and is increasingly being applied to 3MC surveys (Pennell et al., 2017; Słomczyński & Tomescu-Dubrow, 2019; Smith, 2011). Errors in survey estimates comprise variances of estimates (reflecting estimate instability over conceptual replications) and systematic deviations from target parameter values (biases). TSE purports to describe statistical properties of survey estimates incorporating all error sources. That justifies the partitioning of TSE into two components: sampling error and nonsampling error. TSE is also a planning criterion to be used at the survey design stage. The survey designer ideally comes up with a limited number of design alternatives and then selects the design that minimizes TSE. That design should give the highest accuracy for an estimate. However, if there are constraints in terms of costs, timeliness, accessibility, relevance, or something else, the designer must consider the trade-offs to determine if one of the other design options should be chosen instead. This total survey design philosophy was outlined by Hansen, Hurwitz, and Pritzker (1964) and Dalenius (1967). Many other methodologists at the time were also great contributors to the development of the TSE concept.

A popular measure in the TSE framework is the mean squared error (MSE), which is the sum or random errors (variance) and squared systematic errors (bias) across different error sources. The MSE for each individual statistic in a survey is not typically estimated (Groves & Lyberg, 2010; Lyberg & Stukel, 2017; Vehovar et al., 2012), but if we study the components of error and try to estimate the size of those, we can get good guidance concerning where to put our efforts to minimize TSE as much as possible.

In the TSE perspective, there are cost-error tradeoffs, that is, there is tension between reducing these errors and the cost of doing so. As discussed above, Smith (2011) and Pennell et al. (2017) have expanded the traditional TSE framework to include the concept of ‘comparison error’; figures are shown in Appendices 2 and 3, respectively. Słomczyński and Tomescu-Dubrow (2019) highlight the relevance of TSE, along that of fitness for intended use and monitoring survey quality, for *ex-post* harmonization of 3MC survey data; measures for some TSE components for publicly available 3MC survey projects are available in the SDR Database v.1.0, available via Harvard Dataverse (Słomczyński et al., 2017a).

During the last 15 years, TSE framework research has taken a big leap forward. We can identify at least two lines of development. First, the entry of big data, the revival of nonprobability sampling, and access to multiple data sources have resulted in Total Error (TE) frameworks. Another is the development of hybrid (integrated) data sets (Biemer & Amaya, 2020; Groves & Harris-Kojetin, 2017). Still others have developed frameworks for specific big data situations,

where one alternative source is used, such as network data (Marsden, 2011) and other data such as Twitter data, administrative data, and other digital traces of humans (Biemer et al., 2017) (see also Japec et al., 2015). Second, TSE frameworks have been developed for longitudinal surveys (Lynn and Lugtig, 2017). Biemer (2016) has expanded the TSE paradigm into the pillars design, implementation, and evaluation by a system he named ASPIRE that allows survey designers to continuously self-assess their surveys and conduct quality control using statistical process control and quality management principles using critical-to-quality metrics. Finally, Kenett and Shmueli (2014) have suggested a framework they call InfoQ to assess the utility of a data set for achieving a given analysis goal.

These developments are not yet on the agenda for 3MC surveys, but we anticipate that it is just a matter of time. It is hard to imagine that producers and users will abstain from insights based on much more information than what we currently have access to.

TSE is a valuable framework for comparative studies in several ways. First, it provides a blueprint for designing comparative studies. Each component of error can be considered with the object of minimizing error (Fitzgerald, 2015). Second, it is a guide for evaluating error after the surveys have been conducted. One can examine each component and try to assess the level and comparability of the error structures. Third, it can outline a comprehensive and systematic methodological research agenda for studying error and for the design of experiments and other studies to fulfill that agenda. Fourth, it goes beyond examining the separate components of error and provides a conceptual and primarily theoretical framework for combining the individual error components into their overall sum. Fifth, by considering error as an interaction across surveys in multiple study countries, it establishes the basis for a statistical model for the handling of error across surveys.

As discussed by Smith (2011), in the 3MC context, comparability and quality can also be maximized by combining the traditional functional equivalence (FE) approach with the TSE approach (see also Smith (2011; 2019a)). As discussed by Smith (2019a), “concordance of meaning” (Johnson, 1998) is central to the concept of functional equivalence. The FE approach brings a focus to the most important causes of comparison error within the TSE framework and strives to achieve as close a similarity as possible across comparative surveys at both the item and scale levels, by first considering functional equivalence at the item level across matched pairs of questions and then at the scale-level across batteries of items or multi-item scales, which are needed even more in 3MC versus monocultural research (Smith, 2019a). Thus, integrating FE, to achieve functionally equivalent items and scales, along with TSE, to ensure that individual surveys are well designed and well executed to minimize comparison error, can maximize comparability and quality in the 3MC context.

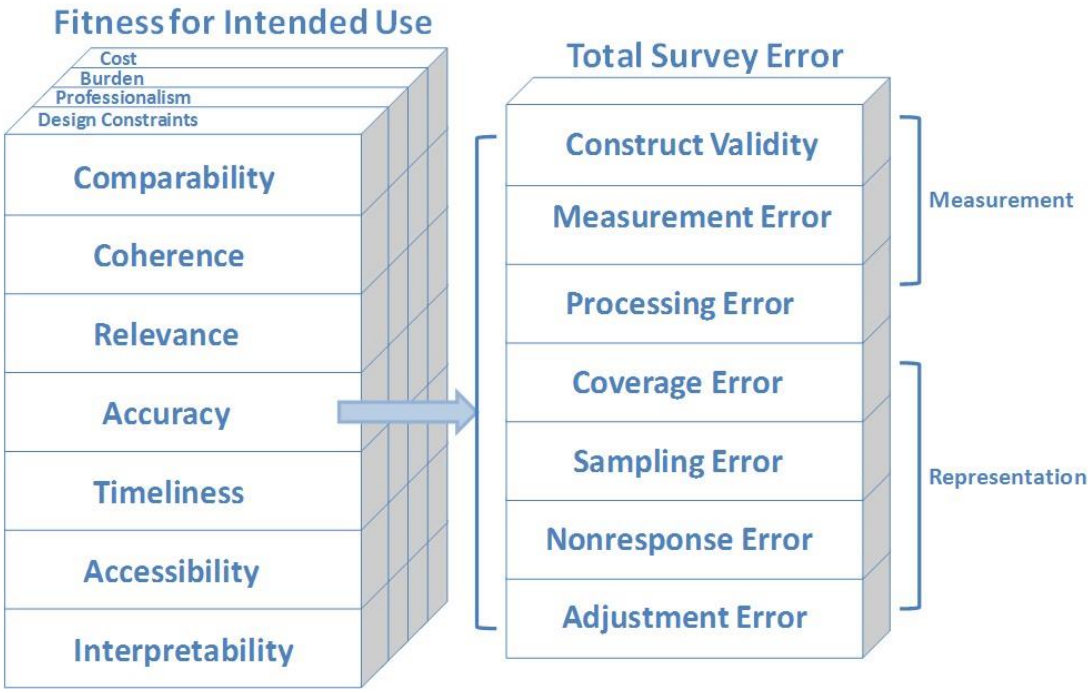
Fitness for intended use

The TSE framework, which has been argued to lack the perspective of users of the data, can be supplemented by fitness for intended use (Biemer & Lyberg, 2003; Gryna & Juran, 2001). Fitness for intended use is multidimensional and focuses on criteria for assessing quality in terms of the degree to which survey data meet user requirements. By focusing on fitness for intended use, study design strives to meet user requirements in terms of survey data accuracy and other

dimensions of quality including comparability, relevance, accuracy, timeliness and punctuality, accessibility, interpretability, and coherence. Fitness for use can be seen as a quality vector consisting of any components that a user is interested in, not just a subset of the ones mentioned here. In this perspective, ensuring quality on one dimension (say, comparability) may conflict with ensuring quality on another dimension (say, timeliness); and there may be tension meeting user needs in terms of both survey error and fitness for use. However, the overall aim is to optimize quality, minimize costs and burden, and recognize and document design constraints at all levels.

This integrated model is visualized in Figure 3, excerpted from Hansen et al. (2016). In this framework, TSE may be viewed as being encompassed by the accuracy dimension in the fitness for intended use model, where *accuracy* refers to whether the data are describing the phenomena that they were designed to measure. That is, are the survey estimates close to the true values of the population parameters they are meant to measure?

Figure 3. Fitness for Intended Use (Quality Dimensions) and Total Survey Error (Accuracy Dimension)



Monitoring survey quality

Monitoring survey quality emphasizes the notion of continuous process improvement (Groves et al., 2009). This framework focuses on quality at three levels: the product, the process, and the organization (Lyberg & Biemer, 2008; Morganstein & Marker 1997). The product quality, as mentioned by Lyberg and Stukel (2010), is the expected quality of survey deliverables, which is often decided by clients and/or other data users. Process quality refers to the quality of the processes that generate the product. One way to monitor and control process quality is to

choose, measure, and analyze process variables, also called paradata or process metrics, relevant to the particular survey (Lyberg & Stukel, 2010). A focus on survey production quality requires the use of standards and the collection of standardized study metadata, question metadata, and process paradata (Couper, 1998), and is operationalized through the quality control process guided by quality planning and quality assurance. The quality control outcome measures are intended to result in a quality profile that can also be used to make recommendations for improvements, and subsequently reflected in future planning. The organizational quality refers to the features that make good processes possible, such as a quality-oriented top management, good user relationships, constancy of purpose, and good competence development programs. There have been efforts to create comprehensive quality frameworks (Ahrendt & MacGoris, 2018; Beullens et al., 2016; Eurostat, 2017; Hansen et al., 2016; International Monetary Fund, 2012; Stoop & Koch, 2013). While these efforts indicate important progress, lacking is an approach that facilitates an assessment of quality in 3MC surveys that provides a comprehensive and explicit focus on the comparative perspective. An investigation of quality in surveys more generally, and 3MC surveys in particular, is often focused primarily on QC protocols during data collection and analyses of output. For example, the European Survey Research Association's (ESRA) 2017 meeting devoted five sessions to "Assessing the Quality of Survey Data", with about 20 papers presented on specific aspects related to quality but few presentations on how quality frameworks might be utilized to evaluate a survey more generally. Similarly, ESRA 2019 had six sessions (about 30 papers) focused on assessing quality, but again the focus was more narrowly on quality in a specific stage of the survey lifecycle. To address this, de Jong and Cibelli Hibben organized a session on approaches to overall quality assessment in 3MC surveys together along with representatives from ESS, Eurofound, PIAAC, and Gallup.

There is not only lack of a consensus on quality standards, but in 3MC surveys it is also very difficult to monitor quality and to develop quality reports that accurately describe the national situation, although there have been some efforts to apply these frameworks to quality assessments. Several 3MC surveys, including ESS, IALS, SHARE, PIAAC, and the Eurofound surveys, have performed both internal quality audits and have commissioned external quality assessments (Börsch-Supan et al., 2008; Gallup Europe, 2010; Vila, Cervera, & Carausu, 2013; Wuyts & Loosveldt, 2019). Such evaluations have largely focused on a review of the processes used and products created to conduct high-quality data collection. For example, previous external assessments of the EQLS focused primarily on the impact of complex sample designs by reporting design effects and standard errors for specific variables, which is a rather narrow focus (Petrakos et al., 2010; Vila et al., 2013). ESS' 2016 quality self-assessment consisted mainly of systematic documentation of the QA/QC processes implemented throughout the survey lifecycle, but an evaluation of utilized systems was absent (Beullens et al., 2016). The ESS assessment was expanded in scope in Round 8 (Wuyts & Loosveldt, 2019).

The most recent external assessment of the EQLS attempted to address the limitations of previous assessments by considering the processes of the 4th EQLS against best practices in the survey research industry, particularly as applied to 3MC contexts. First, a set of 3MC survey best practice guidelines were defined for each of the main stages of the survey lifecycle, considering both the processes of other major 3MC surveys in the European context, including ESS, SHARE, and PIAAC, as well as relevant survey methodology literature, to support the inclusion of each specific standard in the framework. Processes of the EQLS were then considered in relation to

each defined best practice as well as to the process most recently implemented by ESS (de Jong & Cibelli Hibben, 2018). However, this approach has only been used in one study and the extent to which it is applicable to the greater field of 3MC surveys is as yet unexplored.

Quality reviews and evaluation studies are still rare in 3MC surveys. As part of large-scale reprocessing of cross-national survey data and *ex-post* harmonization, the Survey Data Recycling (SDR) Project[7] and its predecessor, the Harmonization Project (Słomczyński, Tomescu-Dubrow, Jenkins, et al., 2016; Tomescu-Dubrow & Słomczyński, 2016) have systematically evaluated differences in survey quality within and between major 3MC projects, including the WVS, ISSP, ESS, EQLS, and Eurobarometer and its regional counterparts, among others. In SDR, survey quality is defined along three dimensions, each informed by TSE, fitness for intended use, and survey quality monitoring, respectively. For each dimension, SDR develops a set of corresponding indicators. These indicators have been applied to public use 3MC data files, to measure variability in survey quality. Ultimately, measures of survey quality inform *ex-post* in SDR, and are stored within the harmonized SDR database (Słomczyński & Tomescu-Dubrow, 2019). Currently, the SDR Team has evaluated 215 data files encompassing a total of 3,485 national surveys fielded between 1966 and 2017 in 169 countries/territories. The first set of evaluations included results from 81 data files containing 1721 national surveys across 142 countries/territories between 1966-2013 (Kołczyńska & Schoene, 2019; Olksyienko, Wysmulek, & Vangeli, 2019; Słomczyński, Powalko, & Krauze, 2017b; Tomescu-Dubrow, Słomczyński, & Kołczyńska, 2017; Zielinksi, Powałko, & Kołczyńska, 2019).

4. Prevailing operational and design challenges

The following section describes key design and operational challenges to quality in 3MC surveys, what we consider current best practice, recent innovations, and future directions. This section follows the major stages or aspects of the survey lifecycle: 4.1 Organizational Structure; 4.2 Sampling; 4.3 Questionnaire Design; 4.4 Translation and Adaptation; 4.5 Questionnaire Pretesting; 4.6 Field Implementation; and 4.7 Documentation.

4.1 Organizational structure

Introduction and key operational and design challenges

The organizational structure considerations and resultant implications for overall study design are particularly relevant to large-scale 3MC survey projects, where, particularly in cross-national studies, there are often multiple stakeholders, study country organizations, and other key actors, requiring substantial coordination in the design and implementation process. The context and current conditions within each country also vary widely, further impacting decision-making. As noted above, a fundamental challenge in 3MC surveys is to determine the optimal balance between local implementation of a design (taking into account the "last mile", or necessary local adaptation) while optimizing comparability across populations. Critical to achieving this goal is a

[7] See https://www.asc.ohio-state.edu/dataharmonization/

comprehensive set of guidelines and requirements that have been developed for each step in the survey lifecycle (Pennell et al., 2017).

Coming up with a set of clear and detailed guidelines and requirements that strikes a balance between standardization and an appropriate level of localization is extremely challenging. As discussed in Section 3, the guidelines and requirements for a study are the building blocks of a QA system, the idea being that by adhering to the requirements, quality is improved. However, having a QA system in no way guarantees quality.

Just as important is a QC system to monitor adherence to the guidelines and requirements and the need to address the underlying causes of deviations from quality standards. As we have noted, 3MC surveys vary in the locus of control and the extent to which it is centralized (all design and operational decisions controlled by a central coordinating team or governing body) or decentralized (each country makes their own operational decisions while adhering to its study design protocols set by the centralized team). While both approaches are in use in 3MC surveys, we argue that a strong centralized infrastructure is needed to maintain adherence to quality requirements.

Current best practices

Generally, the more countries, languages, and populations to be covered in a 3MC survey, the more complex the management task and the need for a strong organizational structure to coordinate all aspects of the design and implementation. Ideally, such a central team is complemented by experts and country representatives that can ensure designs take into account all contexts and local constraints. Such organizational structures can take many forms. Some projects have sought to build a highly centralized organization from which operations are based and where staff is concentrated, whereas others have created more decentralized structures by creating multiple hub institutions that lead surveys across a set of countries (see Appendix B in Cibelli Hibben et al., 2016 for further discussion and examples). No matter the organizational structure, the goal for project leaders must be a deep engagement in the design and operational processes to facilitate input and knowledge transfer of all decisions across the lifecycle that can affect quality both within and across populations.

Regardless of the form, these organizations generally share the following characteristics:

- Comprised of individuals or groups with methodological expertise and experience with 3MC surveys.
- Includes representation from and expertise with the target populations.
- In consultation with all stakeholders, sets design, implementation protocols, and quality standards.
- Monitors and documents all phases of the design and implementation and quality standards, providing support to local teams in real time (or as close to real-time as is possible).
- Conducts methodological studies and facilitates continuous improvement.
- Provides continuous education and support for capacity building.
- Disseminates data and documentation.

- Contributes to both the subject matter and methodological literatures.

These organizations also handle communication among all participating parties to ensure a consistent message. Many have periodic face-to-face meetings supplemented with other modes of frequent communication. These organizations may also handle tender (request for proposal) release and evaluation, contracting, budgeting and distribution of funds. Many are also heavily involved in grant writing and fund raising, given the expense of carrying out these activities.

As 3MC projects develop, leaders must consider what structure best fits the project, recognizing that no two structures may be exactly alike given the differences across projects. However, the development of a strong structure that can address ongoing challenges in research and respond effectively and efficiently, in as close to real time as possible, is critical to producing high quality data.

Recent innovations

While there are many different ways to organize the structure of a 3MC survey, a number of 3MC projects, including the Arab Barometer and the Afrobarometer, have moved from a centralized structure toward a hub structure. Although managed from a central headquarters, the hub structure relies on partners to oversee surveys within sub-regions covered by the project. These regional hubs are trusted local partners that have established high quality research practices. Members from these institutions travel to other countries where survey research traditions may be less well established. The hub institutions lead trainings and remain in-country during initial days of fieldwork to address any problems that may come up.

Given their relative geographic proximity, if problems arise during fieldwork, members of hub institutions can respond to problems without a significant difference in time zones or, if necessary, send a team member to the country to help resolve potential issues. Additionally, as an independent observer, members of the hub institutions can provide feedback about the fieldwork process and provide insight to 3MC project leaders about issues that may have arisen on the ground or improvements that could be made in future survey waves.

The European Commission has funded a number of innovative capacity building projects through its Synergies for Europe's Research Infrastructure in the Social Sciences (SERISS). This funding is aimed at challenges faced in cross-national data collection and breaking down barriers across projects and research infrastructures. Infrastructure projects are aimed at the entire survey lifecycle from study design through data curation. The goal is to 'better equip Europe's social science data infrastructures to play a major role in addressing the key societal challenges facing Europe today and ensure that national and European policymaking is built on a solid base of the highest-quality socio-economic evidence' (SERISS, 2020). To date, this initiative has made a number of strides in funding innovative projects including the enhancement or harmonization of existing tools, developing new software, and providing ongoing training. Examples of these developments are further described in the subsections below. Furthermore, a

related project, the Social Sciences and Humanities Open Cloud (SSHOC), is currently underway aiming at providing data, tools, and training to users of social sciences and humanities data.[8]

Suggested future directions

As 3MC research becomes more complex, challenging, expensive, and yet ever more important in policy formation, funders of such research could learn from the European SERISS and SSHOC examples. Coordination across projects and organizations in the development of new tools and approaches could greatly accelerate methodological developments in 3MC surveys leading to better quality data and increased efficiencies.

4.2 Sampling

Introduction and key operational and design challenges

Probability face-to-face sample designs are currently the most common approach for most high-quality 3MC surveys, although this is likely to change in the future due to increased costs and new data sources. While face-to-face studies are the focus of this section, much of the material is applicable to other modes as well (postal, telephone, and web).[9] In 3MC surveys, complete harmonization of sample designs is not a prerequisite for comparability. In fact, the only stage at which harmonization is necessary and important is in specification of the survey objectives as relating to the sample design and the definition of the target population. The design decisions related to what frame(s) to use, the level of clustering (if any), and stratification variables can and should be optimized on a country-by-country basis (Heeringa & O'Muircheartaigh, 2010), while controlling the process to ensure comparability. This flexibility is important if the objective of the survey is to minimize TSE, both within and across the countries covered. Given this, many operational and design challenges relating to sample design in 3MC surveys are the same as those faced in single-country studies. Yet, there are several issues specific to 3MC surveys which require attention in order to achieve high-quality, comparable data. This section provides an overview of more salient operational and design challenges affecting data quality in 3MC surveys *vis-á-vis* sample design, industry best practices and recent innovations, and future directions for improving quality in 3MC survey sample design.

Countries differ in available frames and thus in how samples can be selected for face-to-face surveys (see Scherpenzeel et al., 2017 for an extensive overview of European countries frames). Several European countries have high-quality person level registers that allow selection of individuals directly, bypassing the household selection stage. Other countries have household or address registers from which a sample can be selected; interviewers then carry out the final stages of selection. In countries without any registers (or where registers are of low quality, out

[8] See https://sshopencloud.eu/

[9] While there has been an increase in the number of nonprobability comparative surveys, these surveys tend to be concentrated in the commercial sector and via opt-in online panels. There are signs, though, that combinations of probability and nonprobability sample designs are gaining ground in other sectors as well. For a comprehensive discussion on sampling approaches for opt-in online panels, readers are directed to the AAPOR report on online surveys (see Baker et al., 2010; see also Chen, Valliant, & Elliott, 2019).

of date, or otherwise inadequate or simply not available), field staff are often used in the selection of households and individuals, although sampling approaches incorporating GIS and other data are gaining more traction, as we discuss in *Recent Innovations* below. As noted above, frames can and will vary in a 3MC survey, and this variability in and of itself does not necessarily challenge data comparability. However, the quality of available frames can differ substantially across countries with regards to coverage, auxiliary information available, and accuracy, leading to significant differences in degree of population representation. Possible sources of variation and errors in registers across countries include the extent of undercoverage of the survey population in the frame, pronounced undercoverage of some sociodemographic groups, inaccuracies in the sampling frame, duplicate registrations, or lack of information to select sampling units with minimum variation of the selection probabilities, such as areas with large apartment blocks.

To complicate matters, sampling terminology can vary across countries. For example, an organization may promise full population coverage based on a specific understanding of the term but then, during the fieldwork period, be unable to interview in the remote areas selected into the sample due to cost and/or accessibility. Measuring the quality of sampling frames across all countries can be challenging due to lack of systematic quality reviews. It leads to lack of information about representation and coverage of the frame and, by implication, of the survey.

Similar issues can arise when selecting stratification variables for use in the sample design and in post-survey calibration weight creation to account for sample design decisions. Such stratification data are most often a result of a prior survey or census in a country, which is subject to data quality issues as well. Additionally, these data are often collected within a single country survey context, and comparable data are not available across all countries in a 3MC project, requiring a harmonization process that itself can also introduce error.

Finally, differences in survey research traditions, survey methodology backgrounds, and variation in socio-political contexts across 3MC study countries can lead to a significant effect on data quality. This can result in misunderstandings regarding sampling methods and even terminology. For example, in a small-scale 3MC study in the Middle East, there was confusion about the term “PPS” (probability proportional to size) which, if not resolved, would have resulted in a significantly different sample design and resultant data in one study country (de Jong & Young-DeMarco, 2016). Definitions of such important sampling terms as *household* and *family member* also vary across countries (Lepkowski, 2005). There may be an uneven level of cooperation among governments when approached with a request for data needed to develop a sampling frame, with variability in cooperation by year as well, further contributing to potential challenges to comparability. Depending on the context, there can be additional challenges inherent to conducting research under repressive socio-political conditions (Tessler, 2011), which may affect the ability to develop an adequate sampling frame.

Current best practices

A number of approaches have been developed to address the key issues relating to sample design in 3MC surveys (Harkness et al., 2010b; Hubbard et al., 2016; Johnson et al., 2019b; Lepkowski, 2005). The following is an overview of those aspects most critical in a 3MC survey, including

standardization of definitions, selection and/or development of sampling frames, the respondent selection within the household, and determinations of sample size and necessary precision. Note that those principles relevant to sample designs in single-country surveys are applicable to 3MC sampling frame development as well.

Target population, survey population and household definitions

One of the key factors when designing a 3MC survey is how to operationalize the design to maximize coverage of the target population and develop a sample design ensuring the survey population covers as much of the target population as is reasonable given cost and operational constraints. It is important to develop a detailed, concise definition of the target and survey populations in order to ensure that each participating country in a 3MC survey collects data from a comparable population (Groves et al., 2009). Differences in the target population may influence estimates of key statistics across countries. Costs and accessibility often mean restrictions in the survey population definition by country. Without a precise definition, countries may differentially exclude or differentially define specific subgroups, such as temporary residents, guest workers, migrants, noncitizens, those speaking rare languages, and institutionalized populations. It necessitates decision-making in each country, which takes into account how different decisions may affect comparability. Of critical importance is careful documentation of the decision-making process and the final sample designs for each country so that comparability can be adequately assessed.

Consideration should also be given to the definition of the household in 3MC studies. A household is a collection of persons who usually reside in the same housing unit (Lepkowski, 2005). Often, the household membership for the purpose of eligibility in a given survey is strictly defined, with definitions including contributing to the common household budget, sleeping under the same roof (for a certain number of nights per week, or the previous night), and eating together. Applying this definition to the diverse living situations all over the world can be difficult and it will be necessary to adapt the household definition to local contexts. Lepkowski (2005: 155) notes that "… in urban slum areas, separate housing units may be difficult to identify when people are living in structures built from recycled or scrap materials." In the Afrobarometer, household membership is defined by the people who at the time of the study eat from the same cooking pot (Afrobarometer Survey, 2014). With regard to the Gallup World Poll, Tortora, Srinivasan, & Esipova (2010) point out that "… polygamy, extended households, and heads of households that rotate among wives' housing units can complicate defining household membership in certain cultures." Again, documentation of such definitions within each country is critical for the assessment of comparability.

Sampling frame assessment

In a probability sample design, all elements in the survey population need to have a known non-zero chance of being selected. A prerequisite for high-quality probability sampling is the use of a sampling frame with very high coverage of the target population in each 3MC survey country. The goal, then, is to select a sampling frame or a set of sampling frames approximating full coverage of the target population to the greatest number of target population elements while containing the fewest number of ineligible elements given survey budget constraints (Groves,

2004). An ideal sampling frame would be fully up-to-date, but real-world limitations often result in at least some level of both undercoverage and overcoverage, duplication, and other inefficiencies.

In a 3MC survey, the best approach is to review available frames in each country, evaluate their accuracy, coverage, and the extent to which data are available for contact and stratification purposes, and assess whether one frame is preferable to another or if a new frame must be developed. The most recent round of the ESS provides a practical example of this process (European Social Survey, 2020). Additionally, as part of one project, the SERISS consortium has produced several documents that assess numerous European individual and address frames on eight key criteria, which should be used when assessing registers (Maineri et al., 2017).[10]

As noted earlier, there can be confusion about what constitutes concepts such as coverage and representation. Ideally, this information should be clear at the tendering (request for proposal) stage and/or implementation of the contract with the data collection organization, well in advance of sample design, to minimize future impact on data quality and comparability. A central coordinating team of a 3MC survey can play a critical role by documenting standards to maximize coverage of the target population prior to sample design development (Heeringa & O'Muircheartaigh, 2010).

Sampling frame development

Concerns over selection bias have led researchers to review their strategies for cross-national studies. For example, in Europe, one of the strategic goals of SERISS is to maximize the use of high-quality registers, ideally, ones that list individuals in all surveyed countries (Maineri et al., 2017). However, not every participating country in a 3MC survey will have a sampling frame that is both accessible and meets the criteria of accuracy and completeness. Where frames are not available or inadequate, common practice is to enumerate (list) households in the selected area units (clusters). This should be done a short period (weeks/months) before the main stage fieldwork and can either take the form of a census of the selected area or a random route with systematic selection of dwellings. However, there are a number of problems with the random route approach. These are further discussed below.

While more costly, enumeration has an advantage compared to a register of being more up-to-date and therefore ensuring all residential households in the selected areas have a chance to be sampled, as well as decreasing the risk involved in giving the interviewers more leeway in the selection of households, as discussed further in the following paragraph. A recent paper by Eckman and Koch (2019), which reviewed the household and individual selection approaches among the countries taking part in Waves 1 to 7 of the ESS, concluded that there was evidence of more bias in samples that gave more control to the interviewer in the selection process and that where more control is given to interviewers, response rates are actually a poor guide to the quality of the fieldwork. They suggest that this is probably due to undocumented substitutions by the interviewers with the intention of maximizing their salaries through completed interviews and/or achieving higher response rates. While the enumeration of households and interviewing

[10] See https://seriss.eu/about-seriss/project-overview/ for more details about the SERISS initiative.

are often done at the same time, this is more likely to lead to the issues Eckman and Koch raise and as such should be avoided or a robust QC process put in place to identify noncompliance.

Where no register is available, one should carefully check that the household selection has been done correctly. A combination of GPS data collected during enumeration for each listed address and satellite/street view imagery, coupled with the address details, allows a reasonable level of central quality control. Additional on-the-ground follow-up checks are also advisable given the limitations of GPS. Such quality control is essential in each country in a 3MC survey in order to achieve data quality and comparability.

Sample size and effective sample size

Another important element in a 3MC sample design is the desired sample size in each country. This is often driven by costs but also requires a sample size that will permit all anticipated analyses, both within and across countries. Depending on the funding structure, countries may have variable sample sizes, affecting relative precision of analyses in cross-national comparisons. Consensus on the analytical objectives of the study is critical to coordinated planning on sample size determination and sample allocation within and across the participating countries. It might also be necessary to boost certain subgroups of the population (e.g., religious or ethnic groups) to allow subgroup analysis and comparisons. The sample design in each country should be designed to maximize efficiency given cost constraints, with a calculation accounting for design effects due to clustering, weighting, and stratification (Heeringa & O'Muircheartaigh, 2010; Cochran, 1977).

Sampling stage determination

Probability samples require probability sampling at all stages of selection. Nearly all established-to-face 3MC surveys employ a multistage clustered sample design where the first stage of selection are geographical areas used as clusters and commonly used as Primary Sampling Units (PSUs). Geographical clustering, while detrimental to the efficiency of the design, is often cost efficient given the travel costs associated with face-to-face interviewing. The ESS, EQLS, EWCS, Afrobarometer, AmericasBarometer, Asian Barometer, EVS, WVS, and PIAAC all use clustered sample designs, with an exception only among a few countries in the ESS and PIAAC that use a simple random sampling selection design. PSUs in each country should be heterogeneous within and homogenous across, as this will minimize the design effects due to clustering. For example, in the UK, postal codes are often preferable to census areas as unlike census areas, postal codes do not define boundaries based on the socio-demographic structure of the household. Availability of electronic boundary maps to facilitate interviewer movement and enable additional quality control is also a consideration when selecting PSUs, although availability of such tools will likely differ across countries and use should be documented.

PPS is the most efficient approach if the same fixed gross sample size in all selected clusters is chosen, as it results in a self-weighting design. It is the preferred method employed by ESS. In some other multinational studies (EWCS, EQLS, GAP) greater focus is placed on achieving a fixed number of interviews in each cluster**,** permitting flexibility in the amount of gross sample issued by cluster. While this will lead to a more geographically balanced net sample, it is a less

efficient design due to the increased variability in the probabilities of selection. In the 3MC context, it is important to recognize the differences among these approaches and apply the selected approach consistently across all study countries (Harter et al., 2010; Heeringa & O'Muircheartaigh, 2010; United Nations, 2005).

In general, for a fixed country-level sample size, the most efficient cluster design is one with a large number of clusters and a small net sample in each. When deciding on the cluster size in each country, as well as the efficiency of the design, one should also account for efficiency of fieldwork, which will vary by country.

Within-household respondent selection

In most face-to-face surveys only one eligible person is selected for interview per household. There are several methods researchers can use to randomly select an individual in the household, including the last/next birthday method, the Rizzo method and the Kish grid. In the 3MC context, household size, perceptions of survey research and associated confidentiality, and other contextual factors relating to respondent cooperation can lead to differential impact of the within-household respondent selection approach. The advantage to the birthday methods is that they avoid the full listing process of household members (Oldendick et al., 1988; Tarnia, Rosa & Scott, 1987). Instead, information on birthdays of household members is used to select the respondent (Salmon & Nichols, 1983). The interviewer first asks for the number of eligible persons in the household, and then asks which person has the next birthday (alternatively: which person had the most recent birthday). The two variants of the birthday method skew respondent selection to those eligible persons born in the months closest to the data collection period of a survey. Birthdays are also not randomly distributed throughout the year and as such these methods do not result in a true random selection. Birthday methods are also susceptible to error if, for example, the household informant does not know the birthday of all household members or the interviewer deliberately substitutes sampled persons.

The Rizzo, Brick, and Park (2004) selection method first asks for the number of adults in the household (*n*). If there is only one adult, the informant is selected. If there is more than one adult in the household, the informant is sampled with a probability equal *1/n*. If the informant is selected, the process ends. If the informant is not selected, the other person is the selected respondent in households with two eligible persons. For households with two or more eligible persons, either the Kish or birthday method is used. The Rizzo et al. approach is particularly advantageous in countries with small average household sizes. See Koch (2019) for a detailed discussion.

The Kish grid is the gold standard for within-household selection in interviewer-administered surveys and involves the listing of all eligible household members (Kish, 1965). The major drawback of the Kish method is that the full listing can be burdensome and time-consuming, and such details as names or initials, gender, and age may be perceived as intrusive (Rizzo et al., 2004). The Kish grid can also be complicated to implement in paper and pencil administration. Concerns are also raised whether the Kish grid might damage the rapport between the interviewer and the informant or respondent and contribute to survey nonresponse, especially in telephone surveys (Gaziano, 2005). In rare cases, a full household listing may not be allowed by

ethics committees, particularly in vulnerable populations. As a consequence, these less invasive methods have been developed. We note, however, these less invasive methods make it more difficult to check that the correct respondent has been selected.

In survey practice, standards are sometimes lowered by restricting the selection of respondents to the persons at home at the time of contact (Holbrook, Krosnick, & Pfent, 2008) or people in the household who are available for an interview on the same day (AfroBarometer Survey, 2014). In the 3MC survey context, such practices at the individual country level may have consequences for data quality and comparability, and only central protocols should be followed.

3MC surveys currently differ in the within-household selection methods they allow or recommend. PIAAC, for instance, requests all countries using household samples to employ the Kish method with a full enumeration of household members (PIAAC, 2014). The use of birthday methods is not allowed. The ESS, EQLS, and EWC allow countries the flexibility to choose between the Kish grid and the last/next birthday, although ESS specifications state a preference for the Kish grid method. The Eurobarometer surveys use the birthday technique (European Union, 2018). The Gallup World Poll offers countries fielding the survey face-to-face (which is the majority of countries) the option of either the last birthday or the Kish method (Tortora et al., 2010).

The Afrobarometer (2014) uses a modified Kish technique and limits the selection to those available for interview on the same day as the initial contact. Interviewers are required to alternate between interviewing a man and interviewing a woman in successive interviews to ensure an equal number of men and women in the sample. The interviewer lists the first names of all household members (in any order) of the respective gender. The interviewer lists the first names of all household members (in any order) of the respective gender. From the list (which is numbered), the interviewer randomly selects the actual person to be interviewed by asking a household member to choose a numbered card from a blind deck of cards. It is recommended that the interviewer ask the male head of household to help select the respondent by drawing a numbered card, as it has been found that this way, they are more likely to understand the randomness procedure and consent to the fieldworker interviewing a young or female household member if selected.

As Koch (2019) shows from his analysis of ESS data, those surveys that involve interviewer selection of respondents are more likely to produce a gender biased sample (with disproportionately more females), compared to samples based on registers. Further, the next (or last) birthday respondent selection method produces greater gender bias than the Kish method. The birthday respondent selection methods have also been criticized for not being truly random. Koch (2019) provides a comprehensive review of these selection methods. Clearly, offering flexibility across countries in a 3MC survey is risky, possibly resulting in unnecessary comparison error.

Central vs. local coordination models

The model used for managing and coordinating a 3MC study will impact the sample design process. In studies such as the Eurobarometer, EWCS, and EQLS, which are all commissioned as

one project by the central body of the European Union, the surveys are conducted by one service provider. Normally in these models the sample design and implementation are done almost completely by a sampling team at the central coordinating center. Countries are consulted on the level of coverage feasible, the frames available, and the level of clustering desired, but the central sampling team makes the final decisions and often takes responsibility for selecting the samples, at least at the areal level. These decisions are almost always made in conjunction with the commissioning organization and the individual countries.

Conversely, the EVS and WVS models provide detailed guidance on the sample design and restrictions on the choice of selection methods allowed but within these parameters the final choice and implementation of that design is up to the individual countries. In contrast, the ESS relies on collaboration between each participating country and the sampling expert panel, with final approval from the latter required before fieldwork commences. Recommended in the Cross-Cultural Survey Guidelines is a central coordination team with support from a group of sampling and methodological experts (Cibelli Hibben et al., 2016). However, key to the success of a central coordination team is its relationship with local providers or country team experts. The relationship must be symbiotic; countries can gain access to technical and expert support, harmonized documents and processes and access to centralized software platforms whilst the central team gains invaluable information on local conditions.

Recent innovations

Recent innovations to address current challenges to probability sampling in 3MC surveys can be summarized into three main areas: population estimates, frame development in the absence of household registers, and strategies for minimizing household and respondent selection bias.

Population estimates

The first challenge and associated innovations concern the availability of up-to-date and accurate areal (geographical) frames for population estimation in the first stage of sampling design. In many low- and middle-income countries, there is limited population data available, and what is available is often out of date. There are often no electronic boundary maps to aid the interviewer in identifying the selected location or the population data is at a very high level of geography making it unsuitable for clustering. In these countries, innovative approaches to sampling using grid-based frames are becoming increasingly popular. Grid-based population estimates were developed to better understand the numbers, characteristics, and locations of human populations, superseding often outdated traditional low-level population data sources that are only updated every decade. They are used in a wide range of fields, including resource allocation, disease burden estimation, and climate change impact assessment. Population density estimates at the 1km grid level (or lower) are calculated utilizing multiple sources of information (census data, satellite imagery, nighttime lights, land use data, the presence of roads, and city and village locations) and state-of-the-art population mapping models.

Table 3 lists several software platforms, all of which provide population estimates at the 1km grid level (or lower) for the world, with the exception of Geostat, which provides a similar product in Europe.

Table 3. Available Software Platforms for Development of Population Estimates	
WorldPop	https://www.worldpop.org/methods
Landscan	https://landscan.ornl.gov/index.php/documentation/#inputData
Socioeconomic Data and Applications Centre	http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
Geostat	https://www.efgs.info/geostat/

Grid-based sampling lends itself well to cluster sampling, often being used as the final areal unit in a multistage sample. For example, the Eurobarometer uses 1km grids developed by Geostat as secondary sampling units in a multistage design. A recent study (Cajka et al., 2018) surveying countries in south and central America, Asia, and Africa also used 1km grids from Landscan as secondary sampling units before going on to select subgrids of variable size in the final areal stage. Grids could also have applications in other types of sampling such as developing an adaptive sampling frame for relatively rare and clustered populations (Peyrard et al., 2013).

There appear to be several benefits over more traditional frames. Grids are very small, making them more suitable where a listing exercise is required. Electronic boundary files make it very easy to develop maps, validate the interviewer's location during fieldwork, and append geographical strata and higher geographical clusters. Up-to-date population data (albeit modelled) enables Probability Proportional to estimated Size (PPeS) sampling. Finally, they are incredibly quick and cheap to build areal sampling frames from, which is ideal if there is limited time to access alternatives. A systematic evaluation of such methods against more traditional frames has not yet been conducted. However, these newer approaches hold much promise, especially where traditional frames are not available.

Algorithmic approaches to frame development in the absence of registers

The second challenge and associated innovations concern developments to address circumstances when population or household registers are deficient or nonexistent, particularly in low- and middle-income countries. In such countries where frame development is required, advances in the availability and resolution of satellite images coupled with the growth of deep learning algorithms has led to the development of models for building detection (Yang, Lunga, & Yuan, 2017) and building footprint segmentation (Bischke et al., 2019). Both techniques aim to identify and demarcate the boundary of buildings using satellite imagery, with the latter allowing a more flexible boundary shape.[11] More recently the Office for the Director for National Intelligence (ODNI) in the U.S. issued the Functional Map of the World (fMoW) Challenge[12], which sought to foster breakthroughs in the automated analysis of overhead imagery by harnessing the collective power of the global data science and machine learning communities. They published one of the largest publicly available satellite-image datasets to date, with more than one million points of interest from around the world. The dataset contains satellite-specific metadata that researchers can exploit to build a competitive algorithm that classifies facility,

[11] See Chew et al. (2018) for an assessment of the suitability of satellite imagery and machine learning techniques to identify buildings for the development of a sampling frame in Nigeria and Guatemala.
[12] https://www.iarpa.gov/challenges/fmow.html

building, and land use. The winning algorithm and others are freely available to download from the topcoder website.[13]

Key advantages of an algorithmic approach over the current 'interviewer listing' are the speed and scale enabled. It also has the potential to improve the quality of the data by removing selection bias at the household level, at least for single family buildings. After the initial outlay to build the models, one would expect to see marked cost savings in fieldwork and increased quality and comparability across countries.

Household and Respondent Selection Bias

The third challenge and associated innovation relates to the bias introduced in the selection of households by interviewers. There is a growing body of evidence in household surveys that the more freedom interviewers have in the selection process, the more the potential for bias, implying that areal frames that rely on random walk or listing and interviewing in one step are likely to lead to larger bias than an individual population register, possibly due to undocumented substitution (Eckman & Koch, 2019; Kohler, 2007; Menold, 2014). These studies conclude that it is preferable, where the option exists, to use individual population registers to minimize selection bias.

There could be a number of reasons why selection bias exists, but it seems reasonable to assume that a motivating factor is the way in which interviewers are remunerated. In many countries, interviewers are paid a fixed amount per completed interview. This type of payment structure is not conducive to behaviors necessary for probability surveys, namely random selection of households and individuals and multiple callback strategies. Short field periods can also induce short-cuts. This is often confounded with the fact that countries less familiar with probability surveys often only have access to areal frames.

Sponsors of 3MC research, when writing tenders and requests for proposals, should promote alternative interviewer payment structures that incentivize the behaviors that increase the likelihood of collecting high-quality data. Also, data collection organizations should experiment with alternative designs rooted in behavioral economics. The setting of targets may also contribute to interviewer deviations from the sampling protocol. Many 3MC survey sponsors recommend targets for the number of conducted interviews by cluster. The use of the word 'target' is misleading and should only be used for convenience samples. Probability sampling requires a minimum level of effort over multiple time periods (during the day, evening, and on weekends) to maximize contact with the selected households/individuals. Setting a target on the number of interviews to complete sends the wrong message. Countries can take it literally and instruct interviewers to achieve the exact target. Given that interviewers vary in their persuasion skills and people's propensity to respond is variable, setting universal targets is not a good strategy. As a matter of fact, achieving precisely the target number of interviews in a moderate to high percentage of assignments is probably a good indicator of selection bias or possible fabrication.

[13] https://community.topcoder.com/longcontest/stats/?module=ViewOverview&rd=16996

When respondent selection is only possible after household selection, Le et al. (2013) suggest an alternative to the Rizzo method, wherein the strategy is more suitable for the larger households that are often present in many settings in a 3MC survey. Their approach alternates between selection methods, conditional on the size of the household and is summarized below in Table 4. To date this method has only been documented in one face-to-face survey in Qatar. However, the results were consistent with the Kish grid approach.

Table 4. Le et al. (2013)'s Respondent Selection Process	
Household size (eligible only)	Selection approach
1	The informant is de facto selected to complete the interview.
2	Randomly select between the informant and the other adult.
3	Randomly select the informant 33% of the time. If the informant is not selected, randomly select between the younger and the older of the other two adults.
4	Randomly select the informant 25% of the time. If the informant is not selected, randomly select between the youngest, the oldest, and the second oldest among the other three adults.
5+	Ask the informant a second question about the number of males in the household. Randomly sample either a male or female. If the number of adults of the sampled gender is less than four, apply the selection method for two- or three-adult households. If the number is four or more, ask the informant to list the names of all adults in the selected gender and randomly choose one.

Finally, we note that the use of computer assisted interviewing greatly facilitates the selection of respondents, both by removing the opportunity for the interviewer to introduce bias into the selection process as well as providing a built-in monitoring system of the process.

Suggested future directions

In this section we outline our expectations for the future direction of sampling in 3MC surveys concerning frame development, assessing availability of registers, and documentation of the sample design process.

The standard random route developed by Noelle (1963) and its variants have been shown in a series of papers by Bauer (2016a) to have the potential to lead to design-induced bias due to unequal selection probabilities. To achieve household selection with equal probability in the absence of an adequate frame, walking instructions should be more flexible and avoid predefined routes that create systematic pathways. In Bauer's latest paper (2016b), he proposes two alternative walks: True Random Routes (TRR) and Street Section Sampling (SSS). In simulations conducted in two towns in Germany, Bauer showed how they produce approximately equal household selection probabilities. SSS requires the identification of a street level frame, which might be problematic, while TRR is very practical to apply and has the potential to be easy for interviewers to understand and follow than the more traditional methods, thereby reducing the potential for errors. The inclusion of a direction grid that interviewers must

follow and use to indicate their routes should also help improve the verification process. These new approaches should be simulated or empirically tested in alternative locations, where street and household locations do not follow such a systematic design or where the names of streets are less clear. Nonetheless, they have the potential to reduce household selection bias if followed correctly (see Appendix 5 for more detail).

Across Europe there is a drive to promote and document the use of individual, household or address registers, and to obtain access to them in countries where previously it has been denied. Recently, the SERISS consortium has produced deliverables on a number of different projects, one of which was to identify, access, and assess registers in each of the member countries. This information can now be found in one central place, making it very easy for survey organizations and their sponsors to access (Maineri et al., 2017).

Outside Europe, although some countries' statistical agencies maintain databases of persons or households for sampling, there is no comprehensive list of those countries, their registers and the quality of them. Those wanting to conduct surveys in these regions using registers need to research the availability of data for sampling frames in each country. For these regions it would be advisable to follow in the footsteps of Europe. 3MC research would greatly benefit from similar efforts in documenting registers in a greater number of countries across the globe. To this end, important questions include: Which countries have a register? What type of register is it? Is the register accessible and has it been used for sampling before? What level of coverage does it have? What information does it contain and how frequently is it updated? Where registers do not exist or are not accessible, further work should be done to understand whether machine learning algorithms can offer viable alternatives, by identifying and mapping buildings to a suitable level of accuracy in the selected areal units (Buskirk, Bear, & Bareham, 2018).

Lastly, comprehensive documentation of the sample design should allow anyone reviewing it to replicate the process should they wish to and to critically assess each stage of selection. There have been great strides in recent years on the level of detail provided in sampling and technical reports in 3MC surveys. However, there is a lot of variation in the quality of this documentation across projects. According to Kołczyńska and Schoene (2019), the best published example across all multinational surveys reviewed is the European Social Survey (ESS). However, even the ESS has been criticized for not providing enough documentation to critically access some elements of the sample design and data collection process (Eckman & Koch, 2019; Menold, 2014). For example, where used, there is limited information on the random walk protocol or how the starting address was selected for random walks. Content standardization across all 3MC surveys with information published and made available, would provide data users with greater awareness of comparability across countries' sample designs and resultant survey data.

4.3 Questionnaire design

Introduction and key operational and design challenges

The goal of questionnaire design for 3MC surveys is to maximize the comparability of survey questions across cultures and languages and reduce measurement error related to question design (Harkness et al., 2016). However, a number of operational and conceptual challenges make

achieving this goal very difficult. We briefly discuss some of the most pressing challenges and also note that, in a 3MC context, questionnaire design cannot be separated from issues concerning translation and adaptation.

This section provides an overview of key operational and design challenges affecting data quality in 3MC surveys with regard to questionnaire design, industry best practices and recent innovations, as well as some suggested future directions toward improving quality in this stage of the survey lifecycle.

While an expansive methodological literature has developed that offers guidance on instrument design in noncomparative contexts, the literature addressing cross-cultural questionnaire design remains relatively small (de Jong et al., 2019; Fitzgerald & Zavala-Rojas, 2020; Harkness et al., 2010a; Wagner, Philip, & Jürges, 2019). Further, any existing design recommendation relating to different types of questions (e.g., behaviors, attitudes, knowledge) or survey mode (e.g., face-to-face, telephone, web), for example, must be carefully considered for appropriateness to the culture and language of each population (Harkness et al., 2010a). For instance, SHARE has made available an overview of all response scales and multiple items used including references that document how they were developed (Mehrbrodt et al., 2017). Due to the array of skills needed to produce a valid and reliable instrument for a 3MC survey, a team approach with subject-area experts, area and cultural specialists, linguistic experts, and survey methodologists contributing at various points during the process is recommended (Harkness et al., 2016; Mohler, 2006). However, it can be challenging to find and effectively manage such a team with the necessary expertise and knowledge. Further, the challenges of documentation, quality assurance, monitoring and assessment for questionnaire design, like other stages in the survey life cycle, are far more complex in the 3MC context.

3MC questionnaire designers also face a number of important issues. Achieving construct validity and measurement equivalence is a salient and persistent challenge. A central question for 3MC survey questionnaire design is whether a given concept exists in a country or cultural context and, if so, how to adapt and operationalize key survey constructs relating to that concept and write questions that are valid and reliable measures of the constructs of interest (Pennell et al., 2017). Further, a good question, with high measurement validity, does not necessarily relate well to the concept it is supposed to measure. Indeed, good questions, in a linguistic sense, may lack construct validity (Billiet, 2016). It is best practice to design questionnaires with multiple indicators to measure a concept, but this increases fieldwork costs and respondent burden, thus, unfortunately, only one question is used in many cases.

In addition to concerns about validity, cultural background has been shown to affect how respondents understand questions and constructs, as well as how they are influenced by the research context. So-called *cultural frames* or *scripts* have been shown to affect the cognitive processes involved in each stage of the survey response process: (1) comprehension; (2) retrieval; (3) judgment and estimation; and (4) response (Schwarz et al., 2010; Uskul & Oyserman, 2006; Uskul, Oyserman, & Schwarz, 2010).[14] The effect of Western European and North American (individualist) and East Asian (collectivist) cultures has received the most

[14] See also Pennell et al. (2017) and Harkness et al. (2014) for overview discussions, and Tourangeau, Rips, & Rasinski, 2000 for a broader discussion on the response process.

attention and is informed by an established conceptual framework and body of experimental evidence (Schwarz et al., 2010). Peytcheva (2020) presents a theoretical framework that maps cognitive mechanisms related to language, such as cultural frame switching and language dependent recall, to the survey response process**,** concluding that these mechanisms "may simultaneously play a role at each step of the response process."

Key features of the measurement context that have received particular attention in the questionnaire design literature are response scales and cultural differences in response styles. Numerous studies have examined cultural differences in acquiescence, extreme, and middle category response styles across different cultural groups in North America (Latino or Hispanic populations and African Americans) and in terms of Hofstede's (2001) cultural dimensions (individualism, collectivism, uncertainty avoidance, power distance, and masculinity) in Europe and across the world (see Yang et al. (2010) for a detailed review and discussion). Response styles may compromise comparisons among cultural and country populations if observed differences in responses at the individual or group level do not reflect true differences on a particular construct. Response scale development and translation are a particularly challenging area because, despite intensive on-going research, it is often difficult to generalize findings from individual studies, research may present contradictory evidence and wording effects often depend on the topic, meaning that clear-cut guidelines on issues related to wording effects in response scales are rare (DeCastellarnau, 2018). For more work on the effect of response styles, see Lee et al. (2019), Liu et al. (2019), and Yan and Hu (2019).

Further development of theory and research on the effect of culture and cognition on survey response is crucial to advancing 3MC questionnaire design. Initial theories have been developed integrating culture in survey response models (Schwarz et al., 2010; Uskul et al., 2010), yet we are still in the early stages and the picture that is emerging is exceedingly complex. Fundamental theoretical debate continues among cultural psychologists about the dimensions of culture, how culture should be conceptualized, and the extent to which culture can be viewed as an explanatory variable. Recent theories view culture as having a dynamic or changing character not restricted by geographical boundaries (Wyer, 2013). Research demonstrates that cultural mindsets can be cued based on the situation producing sharply different "situated" or momentary realities and corresponding perceptions and behaviors (Sorensen & Oyserman, 2009). This indicates that the patterns associated with a particular dimension are not unique to any given country or society but may exist to a greater or lesser extent across societies. If culture is situational or context dependent, it is important to consider how differences in the response process may be influenced in the moment by aspects of the research context (i.e., survey topic, sponsor, preceding questions, question features, the interviewer, the language, and so on), the broader context, or a combination of both. There is, in fact, evidence that contextual cultural cues, such as the language of the interview, can activate specific mindsets and influence survey responses in different ways even for the same individual (Lee & Pérez, 2014; Peytcheva, 2019; Zavala-Rojas, 2018;).

Similarly, important to 3MC questionnaire design are advances in theory and research on the effects of language itself. Language is central to the human experience, and its diversity and range of forms and expressions have produced a wealth of cultural output over the course of history. However, there has been very limited discussion of the role that language may play in

influencing cognition and relevant aspects of the survey response. A recent chapter by Peytcheva (2020) fills this gap by presenting a theoretical framework that maps cognitive mechanisms related to language, such as cultural frame switching and language-dependent recall, to the survey response process**,** concluding that these mechanisms "may simultaneously play a role at each step" of the response process (p. 14).

When designing the questionnaire and other survey materials, researchers must attempt to identify and be informed by ways in which members of different cultures may differ systematically in how questions are understood and answered. Understanding the population of interest and thorough pretesting are essential for the identification of potential problems with design considerations and instruments in order to avoid results plagued by measurement and nonresponse error.

Cross-cultural validity should be established for questionnaires designed to compare data (Fitzgerald & Zavala-Rojas, 2020; Smith, 2004). However, common practice frequently avoids measurement equivalence testing, or equivalence is only tested for a limited selection of items of a questionnaire. Further, pretesting of several alternative measures can be costly in terms of time and economic resources, and sometimes there is little or no compelling evidence to help decide among options (Smith, 2004; Smyth, 2016).

Current best practices

It is important to note that the development of questionnaires for 3MC surveys must begin by following basic best practice recommendations for general questionnaire development (Bradburn, Sudman, & Wansink, 2004; Converse & Presser, 1986; Fowler, 1995; Groves et al., 2009). Also, questionnaire design for 3MC surveys is closely interrelated with translation, adaptation, and pretesting, which are more fully addressed in separate sections in this report.

While relatively limited compared to the literature on questionnaire design in noncomparative contexts, resources exist to guide survey researchers in the development of comparative instruments. For example, Smith (2003, 2004) reviews aspects of general questionnaire design that should be considered when developing instruments for use in multiple languages and provides many references. The Cross-cultural Survey Guidelines (Survey Research Center, 2016) also provide a comprehensive overview. Smith (2015) presents a review of resources, especially those available on the web, which can assist in conducting cross-national survey research (see also Smith, 2019b). Revilla, Zavala-Rojas, and Saris (2016) show how to use evidence from experimental research to design questionnaires aimed for multiple populations. Harkness, van de Vijver, and Johnson (2003) present advantages and disadvantages of major comparative design models and a general framework for design decisions. These include the sequential approach, where the source questionnaire is developed by a small group and then translated, the parallel approach, where the source questionnaire is developed so that it is appropriate for all target cultures, and the simultaneous (i.e., decentering) approach, where questionnaires are developed in multiple contexts, with a common core of concepts implemented in an approach specific to each country.[15] Harkness et al. (2010) further develops this framework and also address basic

[15] The framework outlined in Harkness et al. (2003) has been largely supplanted by further work by Harkness et al. (2010).

considerations for comparative questionnaire design. The three basic approaches involve asking the same questions and translating (ASQT), asking different questions (ADQ), usually to adapt to new cultural, social or other needs, or using a mixed approach that combines ASQT and ADQ. Choosing a questionnaire design strategy depends on the survey's research questions and the cultural portability of the concepts to be measured (see Harkness (2008), Harkness et al. (2010a) and Harkness et al. (2016) for details and a discussion of pros and cons of these approaches).

Regardless of the design strategy chosen, a conceptually oriented design is considered best practice. This refers to the theoretical definition of the concepts to be measured and the operationalization of those concepts into latent variables, indicators and survey items. Careful choice of linguistic elements for questions and response scales to minimize method effects is essential. Ideally, more than one indicator should measure a concept. When concepts are measured with more than one question, estimation (and statistical correction) of the errors of measurement is possible (Rammstedt et al., 2015).

As discussed above, questionnaire design teams should be multidisciplinary, including survey experts and theme or subject matter experts. Design teams should receive multicultural input throughout the process (multicultural design teams, input from participating countries, comparative pretesting, and so on) to evaluate cross-cultural challenges.

Lastly, some form of questionnaire pretesting before fieldwork is common current best practice. Testing a questionnaire before the main fieldwork to decide among the wording options is important to achieve reliable measures. A combination of qualitative and quantitative methods should be implemented to triangulate information on the performance of the questions (Fitzgerald & Zavala-Rojas, 2020), for instance, split ballot (quasi) experiments and cognitive interviewing (see the Pretesting section below for further discussion).

Recent innovations

Anticipating problems with survey items before they are administered to respondents has been an intensive area of research in recent decades. Translatability assessment (TA) or advance translation (AT) are relatively recent approaches that have been developed to detect cultural/linguistic issues and check whether the text will be easy to translate and identify potential translation challenges at the early stages of 3MC questionnaire design. As a result, the source instrument may be revised or specifically annotated for translation. Both approaches, TA and AT, have been found to be effective at enhancing the translatability and the cultural adaptability of source questionnaires for translation into multiple target languages (Acquadro et al., 2018; Dorer, Forthcoming). While TA and AT share many features, there is one crucial difference. TA relies on the work of individual translators to translate the source questionnaire into several languages and/or review the source text in terms of translatability. Comments on each individual language from these translators, who typically have had training specifically as translators, are then merged by a project manager into one common file. AT, on the other hand, involves translations that follow a team approach: interdisciplinary teams composed of translators and survey researchers apply a multi-step translation approach and discuss both the translators' and the survey researchers' comments in a review session. In a final step, all comments are not only copied together by one person but discussed and agreed by all actors

participating in a review discussion. In this way, some argue that AT offers a clear advantage for survey questionnaire development (Dorer, forthcoming). However, one drawback of the AT approach is that it tends to be more costly because a team of at least three experts needs to be employed for each language assessed.

Translation annotations and notes often resulting from the above procedures, should supplement a source questionnaire where this is deemed useful (Behr & Scholz, 2011; Dept, Ferrari, & Halleux, 2017). These annotations include additional information for translators, such as the intended meaning of a key term, a cultural adaptation instruction or a pointer to a particular design decision.

Several (software) tools that have been developed to aid questionnaire designers can also facilitate 3MC questionnaire design. The Question Appraisal System, QAS-99 guides users to check whether a systematic evaluation of potential cognitive problems in the draft texts exist (Willis & Lessler, 1999). The QAS-04 is an attempt to adapt the QAS to the cross-cultural context (Dean et al., 2007). The Question Understanding Aid, QUAID is an online tool that assists questionnaire designers in identifying problems with the wording, syntax, and semantics of survey items (Graesser et al., 2006). The Survey Quality Predictor, SQP is an online software that allows prediction of measurement errors, providing suggestions of desirable questions' features based on a meta-analysis of hundreds of experiments testing question formulations (Saris et al., 2011). Zavala-Rojas, Saris, and Gallhofer (2019) discuss strategies for preventing differences in the measurement properties of translated survey items using the Survey Quality Predictor (SQP) system.[16]

A software tool has also been developed by GESIS in Germany to provide context-sensitive measurement and simple harmonization of educational attainment, a key background variable that can be complicated to measure comparatively due to the different educational systems in different countries, increasing differentiation of educational systems, and increasing education and work-related migration. The Computer-Assisted Measurement and Coding of Educational Qualifications in Surveys (CAMCES) tool measures educational qualifications in computer-assisted surveys, based on: 1) an international database of educational qualifications; 2) optimized questionnaire instruments; and 3) an interface to directly access the database for use in computer-assisted surveys. It is free to use, with support provided for academic surveys.[17]

Specialized software for questionnaire design management and documentation is another promising area. Questionnaire design in comparative survey research is an elaborate process; questionnaire iterations and evidence from pretesting accumulates rapidly in large-scale survey projects, word processors are not enough, electronic management systems are necessary for processing and summarizing pretesting findings, and version tracking can be complicated. The Questionnaire Design and Documentation Tool (QDDT), for example, is a free web-based software tool developed with the ESS based on an earlier paper template approach (Fitzgerald, 2015), as its primary use-case for documenting and managing information on the complex process of designing a cross-national survey questionnaire. Based on the Data Documentation

[16] See http://sqp.upf.edu/

[17] See www.surveycodings.org/levels-education and Schneider et al. (2018) for more information.

Initiative (DDI), QDDT captures and displays the development history of survey items (Orten, Norland, & Butt, 2018).

A number of important innovations have also emerged in the area of research and methods for examining construct and measurement comparability, many of which are featured in the recent 3MC monograph, where de Jong et al. (2019) provide an overview. These include advances in psychometrics that have allowed survey researchers to test measurement equivalence empirically for questions that meet certain requirements; strategies for detecting and addressing differences in question sensitivity in a comparative context; the re-evaluation of a series of classic split-ballot questionnaire experiments previously conducted in monocultural settings in an online multinational study (Silber et al. 2019); the use of anchoring vignettes including an innovative sensitivity analysis; cognitive interview methods for evaluating question comparability, and behavior coding as a method for comparing the cognitive processing of survey questions across cultural groups. For a more detailed overview, see de Jong et al. (2019).

Suggested future directions

Awareness needs to be raised about the challenges of 3MC questionnaire design, current best practices, and the importance of investing sufficiently in the questionnaire design process to produce instruments that are as valid, reliable, and comparable across populations as possible.

As noted above, further research and the development of theory are needed to integrate culture into the survey response process. If culture is situational or context dependent as current cultural psychology theory posits, it is important to consider how differences in the response process may be influenced in the moment by aspects of the research context (e.g., survey topic, sponsor, preceding questions, question features, the interviewer, and so on), the broader cultural context, or a combination of both. Similarly, research on the understanding that language itself plays in the survey response process is necessary, both through and separate from culture. Ji et al. (2004) discussed the tension between culture and language nearly two decades ago, but there has been little advance on the topic since then. Future research should go beyond global country comparisons and take into account the complex interplay among factors stemming from the individual, the survey context and the broader context, and examine a broader range of cultural dimensions. Further interdisciplinary collaborations between survey methodologists and cultural psychologists and others specializing in cross-cultural research could help strengthen the theoretical foundations of 3MC survey research.

Many 3MC surveys have started to explore mixed-mode designs for potential costs savings, among other reasons, and mixed mode approaches are likely to feature prominently in the changing survey landscape (Lyberg et al., 2019). Results of mixed-mode experiments by the ESS showed there to be more cons than pros in mixed-mode approaches to data collection across countries, leading the ESS to reject such a switch (Villar & Fitzgerald, 2017). However, while it is still feasible for the ESS to maintain a single-mode framework mitigating mode effects, a mixed-mode approach in 3MC surveys is not uncommon and is at times inevitable due to differences in survey traditions and survey climate, literacy, availability of registers and sampling frames, differences in Internet and telephone penetration, and available fieldwork budgets. Optimal mixed-mode design decisions, through a unified mode design, for example, can

help mitigate mode effects; however, when different modes are added to a design that already includes different countries and languages, measurement problems are likely to increase and rules for equivalent questionnaire design are of the utmost importance (de Leeuw, Suzer-Gurtekin, & Hox (2019) provide a review for 3MC surveys). While progress has been made on the development of mechanisms to disentangle 'selection effects' from 'measurement effects' (Vannieuwenhuyze & Loosveldt, 2013; Vannieuwenhuyze et al., 2010; 2014) and implementing approaches to collect additional data to adjust for both mode selection effects and mode measurement effects at the analysis stage, further research is needed to improve the cost-effectiveness and practicality of these techniques (Villar & Fitzgerald, 2017).

Progress in 3MC questionnaire design would be facilitated by a central resource or database with tested questions and information on what has been found to work and not work in comparative questionnaire design (e.g., problematic terms, linguistic structures, indicators, and lessons learned from major studies).

4.4 Translation and adaptation

Introduction and key operational and design challenges

Translation is possibly the step that most easily comes to mind when thinking of the differences encountered in 3MC survey operations. When referring to translation, we also include adaptations, i.e., intended deviations from source items so that they are relevant and appropriate for a new cultural context. It is important to note that questionnaire translation is crucial for the comparability of the resulting data. However, good and comparable translation is not sufficient in itself. As the preceding section on questionnaire design illustrates, a careful source questionnaire design process that focuses on the portability of questions across countries and languages in terms of relevance, validity, and translatability is essential at the outset. Thus, we presuppose in the following discussion that a suitable source questionnaire is available for the translation task, ideally annotated with information for translators where this is deemed useful. We also do not cover the pretesting of translations here; this topic is covered in the next section. Furthermore, we note that the following general considerations apply to translation regardless of survey mode, even though there may be particularities linked to individual modes, such as placeholders and their linguistic challenges in computer-assisted surveys (Behr, forthcoming; Pan, Sha, & Park, 2019), or particularities linked to mixed-mode surveys where language formality may differ between self-administration and interviewer administration. This text is intentionally general and brief. Step-by-step translation guidelines can be found, for example, in the Cross-Cultural Survey Guidelines (Mohler et al., 2016). See also Smith (2008) for additional thoughts on translation, including possibilities for quantification. The following covers operational and design challenges affecting translation data quality in 3MC surveys, industry best practices and recent innovations, and future directions.[18]

[18] The translation/adaptation of other survey materials, such as recruitment letters, consent forms, and web sites is not covered here since the comprehensive translation processes described in this report typically do not apply to these text types. Nevertheless, for budgeting and/or planning other types of quality control, these further translation needs should be taken into account.

While translation may come easily to mind as an essential step in the process of collecting comparable survey data, important misconceptions mean that the design and operational challenges related to translation and adaptation are often not well understood or given adequate consideration in many 3MC survey efforts. First, translation is often misjudged to be a quick and easy process that does not require much skill, time and resources. This misconception hinges on the fact that translation is often regarded as a mere mechanical, almost automatic word replacement process (Colina et al., 2017; Gambier, 2016). However, this is an outdated view of translation work both from our knowledge of the translation literature and the translation industry. For over four decades, scholars in translation studies have developed modern – *functionalist* – theories and frameworks as to what translation actually is and how to approach it (Calvo, 2018; Nord, 2014), and these theories and frameworks are consistent with what is covered in internationally accepted translation standards, in particular the ISO standard 17100:215, "Translation services -- Requirements for translation services" (International Organization for Standardization, 2015a). The one and only translation that naturally – and closely – flows from the source questionnaire does not exist; a good translation requires taking into account many factors, which differ from project to project and which include in particular the purpose and use of the translation, the target group, typical text type and domain features, and other project-specific requirements.

A second misconception is that anybody who can speak two languages (in one way or another) is able to produce a good translation. The thinking behind this is that if you know two languages you will surely be able to replace words from one language with words from another language, which is then linked again to the first misconception. Yet, translation studies show that translation novices tend to translate on a word-by-word basis rather than taking into account important context or project information (purpose of translation, target group, and so on; Göpferich & Jääskeläinen, 2009; Jääskeläinen, 2010; Shreve, 2002). Furthermore, translation requires more than "mere" language competence and, in fact, draws on a number of different competencies, including translation competence as such (translating in line with project specifications), linguistic and textual competence in the source and target language, cultural knowledge, competence in the substantive domain as well as in research, information acquisition and processing, and eventually in information and communications technology (ICT). For questionnaire translation, specifically, which places heavy demands on the clarity and simplicity of wording and syntax, a very good feeling for the target language is crucial as is the ability to write for a general audience of all educational levels. Of particular importance is also knowledge of questionnaire design (Behr, 2018).

An additional misconception is that adequate translation can be accomplished "on the fly" – an ad hoc approach sometimes employed if only a small number of respondents are expected to need a specific language version, but not enough to justify production of a written translation. Beyond knowing that these translations are made orally, little can be said about the approach taken in any specific case (Harkness & Schoua-Glusberg, 1998). While the appeal of such a strategy in countries with upwards of eight or more languages is clear, the lack of standardization can lead to significant measurement error and uncertainty in the ability to compare data across populations, both within and across study sites.

The above misconceptions often lead to simplistic translation and checking procedures in 3MC surveys, and in particular to the over-reliance on back translation as a quality control method (Bolaños-Medina & González-Ruiz, 2012; Chidlow, Plakoyiannaki, & Welch, 2014; Colina et al., 2017). Back translation (Brislin, 1970)[19] includes, in its simplest form, the translation of the original translation back into the original language and the subsequent comparison of the two questionnaire versions in the original language. Discrepancies discovered may give rise to changes in the original translation. While back translation is certainly able to pick up some translation problems, there are limitations to the method. Most importantly, back translation cannot assure suitability for a new context when the instrument should have been culturally adapted. Furthermore, since in a simple back translation design there is no systematic review of the actual translation involved, complex or difficult wording and syntax, inappropriate register for the target group, and incorrect spelling often cannot be identified. In addition, the back translator may iron out problems that exist in the original translation (they are no longer visible in the back translation) or even introduce mistakes, which either leads to a false sense of security or to a false alarm (Behr, 2017; Colina et al., 2017; Douglas & Craig, 2007). For the above reasons, back translation should not be used, at least not as a sole quality check. For documentation purposes, however, some authors see its value (Son, 2018).

Current best practices

Addressing the limitations of back translations and, in general, leading the field forward, 3MC survey methodologists have developed alternative methods and best practices that are believed to produce higher quality translations.[20] The underlying principles with these alternative methods is that quality of the translation requires in-depth assessment of the translation itself and that quality is based on a multistep process that requires people with various skillsets who implement this process. The model TRAPD – Translation, Review, Adjudication, Pretest, and Documentation has been one of the key methodological drivers in this regard. It conceptualizes the activities that should be used when translating survey questionnaires. The model foresees team-based approaches to translation, which date back to the 1960s when they were used in Bible translation efforts (Nida & Taber, 1969), along with the major innovation of integrating pretesting and documentation in the process. Even though empirical research testing the success or usefulness of TRAPD – or its variations – is lacking, as we discuss further below, consensus is growing across several disciplines around a multistep translation and review process, bringing in different skillsets, using some form of pretesting, and documenting the entire process (Acquadro et al., 2008; International Test Commission, 2017; Wild et al., 2005)[21] as current best practice for translation quality and transparency. First developed for the (ESS) (Harkness, 2003; Harkness, Pennell, & Schoua-Glusberg, 2004), TRAPD is now widely used in the global survey research

[19] Brislin (1970) himself cautioned against the exclusive use of back translation as a quality control procedure. However, this qualification seems to be widely ignored when Brislin is cited as a proponent of back translation.

[20] Back translation may still be conducted alongside these best practice methods: In other disciplines, the back translation step is not discarded altogether but may be implemented alongside best practice procedures (Acquadro et al. (2008) and Wild et al. (2005); but also see McKenna and Doward (2005) for a critical review of back translation).

[21] Note that in other disciplines, in particular the health sciences and psychology, psychometric assessment is part of the overall process of translating and adapting an instrument (e.g., International Test Commission, 2017).

community, although not always labeled as such or implemented in its complete form. An overview of the TRAPD model, as originally conceived for the ESS, is presented in Table 5.

Table 5. The TRAPD Model - Translation, Review, Adjudication, Pretest, and Adjudication[22]

Step	Description	Rationale	Staff
T	**Translation**: The questionnaire is translated by two translators independently from each other, resulting in two versions of the questionnaire.	Two versions help to spot meaning differences, idiosyncratic wording, and also mere oversight. They also provide different stylistic variants to choose from or work with.	**Translators**: Skilled translation practitioners, i.e., persons with translation experience, and/or who have been trained on the particularities of questionnaire translation.
R	**Review**: In a review meeting, a reviewer and the translators jointly reconcile the two versions into one.	Different expertise (translation as well as questionnaire design and topic) is regarded essential for producing a high-quality translation, and so is a direct exchange rather than a step-wise consecutive review by individual persons.	**Translators**: Those having produced the translations at step T. **Reviewer**: Survey researchers or study managers bringing in additional competences on questionnaire design, survey research, and the topic.
A	**Adjudication**: In the adjudication phase, any pending issues from the review are taken care of and the translation is signed off for the next step.	During the review meeting, questions may come up on the measurement goal of some items; also, a questionnaire translation requires careful copy-editing looking out for completeness, consistency, and overall coherence.	**Adjudicator:** The adjudicator may be the same person as the reviewer, depending on how responsibilities are divided in a team. This is the person assigned for the ultimate decision-making.
P	**Pretest**: The translated questionnaire is pretested among the target population.	Translated questionnaires should be as thoroughly tested as any other questionnaire. Different qualitative or quantitative testing procedures are available.	
D	**Documentation**: Particular decisions and adaptations during the entire process are documented; documentation also involves describing the different processes that have been employed (and by whom) to arrive at the final translation.	Documentation provides useful information both for internal monitoring and for external data users.	

[22] Table 5 presents the TRAPD model as originally implemented in the ESS.

It needs to be acknowledged that TRAPD, as described above, is expensive and labor intensive. Hence, a number of different variations of TRAPD have emerged, typically in response to practical constraints in terms of cost, time, and feasibility. TRAPD variations often reflect how the translation (T) and review (R) steps are carried out. For example, in what is called a split or modified committee approach to translation, two or sometimes even three translators translate a questionnaire (Harkness & Schoua-Glusberg, 1998; Schoua-Glusberg 1992) but instead of producing two or three full versions of a questionnaire translation, the instrument is split up among translators and each one translates parts of every topic or section. It is thought that the translators are able to gain sufficient familiarity with the questionnaire to usefully contribute to the team discussion by having worked on each section, at least to some extent. At the same time, the split approach saves time and money. Another variation involves having one translation produced by a single translator. This translation is then discussed and further refined by an interdisciplinary team, which brings translation as well as survey and topic expertise to the table but may exclude the actual translator. The team composition at the second phase is seen as compensation for the "missing" second translation and for a potentially missing pretest (Goerman, Meyers, & García Trejo, 2018). Yet another variation consists of double translation and subsequent reconciliation by a single person, with or without further discussion. Regardless of whether these or even further variations meet the original ideas of TRAPD or not, it is important to consider the advantages and disadvantages of each method prior to its use. Ultimately, it should be noted that the persons employed and their various skills will have a large impact on the success of a method.

Regardless of the exact approach chosen, it is best practice for the translation and adaptation process to involve the following: First, translation and the subsequent checking processes require time and financial resources. This needs to be considered by those planning and budgeting a study. Second, translators and reviewers need to be briefed (if not trained) on the requirements on the translation (purpose, target group, mode, and other requirements), so that they can make translation decisions in line with the study goal (Calvo, 2018). This necessary step of providing project specifications is also stressed in the ISO standard on translation (International Organization for Standardization, 2015a). Third, when it comes to reconciling two translations (in the case of double translation), it is essential that this should never be limited to just selecting one or the other translation, or merely combining them. Reconciliation requires in-depth reworking and may even lead to completely new versions in the light of previous decisions. There has been some concern as to the effect of undesired group processes, or withholding of criticism, in team or committee approaches (Brislin, 1980; Koller et al., 2012). A few strategies can mitigate these concerns, amongst which is a transparent description of translation processes prior to translator selection, a skilled reviewer leading the discussion and ensuring balanced participation, and sufficient time for discussions (Behr & Shishido, 2016). Further, after having implemented a multistep translation and review process, translations should be tested among the target population, e.g., qualitatively in cognitive interviews or quantitatively in pilot studies (discussed further below). Finally, the translation process, including particular and/or difficult decisions, should be documented for both internal and external uses.

Sociolinguistics examines how language is used in its social contexts and their functional aspects. It provides a theoretical framework for survey translation and complements the procedures described in the TRAPD model and its variations. According to Pan, Sha, and Park (2019), functional equivalence or comparability in survey translation should be achieved in a sociolinguistic framework. To do this, translations must be considered at different levels of language use. For example, it can be relatively straightforward for a translation to satisfy the lexical and syntactic requirements of the target language. However, it is challenging for the source text to achieve the equivalent communicative effect, referred to as the pragmatic level. A translation is appropriate at the pragmatic level when it is rendered in a linguistically accurate structure of the target language ("linguistic rules" are followed) while incorporating the social practices and cultural norms of the target culture.

Recent innovations

Regardless of the particular translation method employed, the translation field is currently undergoing a rapid technical development, which consists of the increased employment of translation management platforms and computer-aided translation tools (either those used in the translation industry or adapted or newly built ones) and, hence, of the greater use of the potential that is linked to such platforms and tools (version control, leverage of existing translations, terminology data bases, automatic checking procedures, facilitated documentation, and so on). The OECD studies PISA and PIAAC (with their use of the open-source translation tools OLT and, more recently, OmegaT) or the European study programs SHARE, ESS, and EVS (with their use of the proprietary Translation Management Tool (TMT)) are cases in point. Innovation has also meant an increased research interest in how to incorporate technologies developed in consolidated areas of language disciplines into the translation of survey questionnaires. In Europe, the SERISS consortium supported research projects in several areas of innovation, for instance, how to incorporate tools from corpus linguistics and computational linguistics into survey translation, quality assessment of thesaurus keywords of survey items, and comparative testing of questionnaire translation approaches.[23]

Looking ahead, the role of machine translation – however small or large it may be – will certainly also need to be investigated in the near future, given that neural machine translation is significantly improving the quality of machine translation output – and progress is steep in this field. For instance, in the Social Sciences and Humanities Open Cloud (SSHOC) project, European survey infrastructures are experimenting with the use of machine translation in different translation settings.[24]

Suggested future directions

Going forward, while debate about the method of translation continues and the supporting technology evolves, the matter at the heart of it all – the nature of the relationship between the target questionnaire and its source – remains unclear. Remaining close to the source questionnaire in semantic and measurement terms, while adhering to target language needs, is

[23] See https://seriss.eu/about-seriss/work-packages/wp3-maximising-equivalence-through-translation/; outputs can be retrieved from https://seriss.eu/resources/deliverables/.
[24] See https://sshopencloud.eu/.

still the gold standard in 3MC survey research that follows the ASQ (ask-the-same-question) approach. It is feared that deviations result in measurement artefacts. We need to learn more about the impact of deviations on the comparability of data (e.g., what it means to change the semantics in response scales or what it means to delete (or add) information that is (not) existing in the source). Along with this, we have to better understand the notion of a deviation, because formal deviations in the sense of a lack of "formal equivalence" may in fact lead to "dynamic equivalence", resulting in the message of the source text being accurately transposed to the target language (Nida (1964); see also Kleiner, Pan, & Bouic, (2009); Zavala-Rojas et al. (2018)). A better understanding of the impact of different translation options on the resulting data will strengthen the position of translation in the survey lifecycle. It will contribute to a realistic understanding of what good translation entails and it will provide reasons for carefully designing, implementing, and documenting translation and assessment processes. In sum, the future will need to see more qualitative and quantitative (experimental and evaluation) studies focusing on translation quality assessment.

As 3MC surveys increase in scope to include more countries and languages, equivalence *across* language versions – beyond the 1-to-1 relationship between the target and the source – is often at risk. This threatens the comparability of any combination of questionnaires in subsequent data analyses. "Harmonization" procedures both during questionnaire design (translation annotations; see discussion of advance translation and translatability assessment in the questionnaire design section above) and during translation (webinars/joint meetings, query lists) are possible solutions (Behr & Shishido, 2016) to render translation decisions that are consistent across countries and to effectively manage and document the process. Furthermore, external quality control delivered in a consistent manner can be seen as a way to monitor the output and increase equivalence across a multitude of versions.

As noted above, to date, only few empirical studies have examined the success or usefulness of TRAPD, its variations, or other methods in comparison. Behr (2009) looked into how discussions evolved during a team review meeting and what was (or was not) discussed. Schoua-Glusberg (2004) examined committee discussions that revealed how decisions are made in the absence of translation specifications. Scholars from the health sciences (Epstein et al., 2015; Hagell et al., 2010) have produced evidence supporting team approaches (expert committees or dual panels) as opposed to back translation approaches. Further research is urgently needed for empirically backing or rejecting individual translation and assessment methods.

Even if researchers are well aware of and wish to implement TRAPD or variations thereof, the implementation of such procedures can be a challenge due to the shortage of translators experienced in translating questionnaires, of skilled reviewers, and also of organizations that are capable of conducting the full process in a multitude of languages. Capacity building is certainly needed in this regard. AAPOR and ESRA, amongst others, regularly offer webinars and courses on the topic of questionnaire translation, so that at least survey researchers can become aware of best practices in the field. Furthermore, major publications and guidelines are available online, such as the Cross-Cultural Survey Guidelines (Survey Research Center, 2016) and related short courses, the translation guidelines from the ESS (European Social Survey, 2018b) and those from the US Census Bureau (Pan & de la Puente, 2005).

4.5 Questionnaire pretesting

Introduction and key operational and design challenges

Pretesting techniques typically used in single population surveys can also effectively be applied in 3MC surveys (Caspar et al., 2016). These include qualitative methods such as cognitive interviews, focus groups, expert reviews, and field tests (Presser et al., 2004). Where sample sizes are sufficient, quantitative techniques such as latent class analysis, item response theory, Multi-trait-Multi-method (MTMM) experiments and other approaches are also increasingly applied (Saris & Gallhofer, 2014). Discussion of these latter techniques can be found in Harkness et al. (2010a) and Caspar et al. (2016), and examples of approaches that combine quantitative and qualitative approaches can be found in Benitez and Padilla (2014) and Fitzgerald and Zavala-Rojas (2020). The following section focuses on qualitative approaches to pretesting and provides an overview of operational and design challenges affecting data quality, industry best practices and recent innovations, and future directions.

Different types of pretesting techniques tend to yield different types of results and no one technique can offer a comprehensive set of findings about the quality of or potential problems with a questionnaire. It is therefore ideal to combine pretesting techniques and/or post-survey analysis methods in a way that takes advantage of the strengths and minimizes the weaknesses of each method. These techniques can be combined to provide a comprehensive, and ideally iterative, pretesting design. For further discussion, see Caspar et al. (2016), Fitzgerald, Winstone, and Presage (2014), Smith (2019b).

Various pretesting techniques can be applied both before and after survey questions have been drafted or translated. For example, focus groups can be useful in formulating questions whereas other approaches such as cognitive interviews and expert review can test the source questions and translations. The results of pretesting help researchers improve question wording and translation, determine the degree of confidence they have in the comparability of the questions, and evaluate whether the questionnaire is ready for data collection. Therefore, pretesting is essential in questionnaire design and serves as a key step in the overall 3MC survey lifecycle. While pretesting plays a critical role in identifying and potentially reducing measurement error that harms statistical estimates at the population level and thus endangers comparability across populations (Caspar et al., 2016), few standards currently exist for the type, combination, or amount of pretesting that should be done in 3MC surveys (de Jong & Cibelli Hibben, 2018).

3MC studies face both methodological and practical challenges in designing and implementing pretests. For example, carrying out pretesting such as cognitive interviewing in 3MC studies typically involves coordination with multiple partners in various locations, across different time zones and a variety of languages (Miller, 2019). The availability of trained staff and respondents' familiarity with this research approach also often varies considerably. As Miller (2019) notes, in some cases, those who are conducting the interviews may have never before conducted a cognitive interview or participated in qualitative research of any kind. When working with multiple languages, it may be necessary to transcribe and translate interviews (or provide detailed summary notes, at a minimum) into a common language so that they can be reviewed to ensure quality and comparability of interviewing approaches. Further, cognitive interview results

must themselves be translated into a common language, if these are to be discussed by researchers who do not speak the target languages (Willis & Miller, 2011).

As Pennell et al. (2017) note, even when a pretest is carefully carried out, interpreting and integrating results may prove difficult. Results may highlight problems with survey questions but solutions may be less obvious or identified solutions may solve a problem in one context but create harmonization problems with the measure in other contexts, particularly for items that may have been used extensively in other contexts or are needed for comparability across time. Furthermore, changes made to a questionnaire based on cognitive interviewing or other pretest outcomes may necessitate a subsequent round of testing.

Another challenge is that some studies have shown that different cultural and linguistic groups may respond differently to pretesting methods. For example, Pan et al. (2010) found in a study of monolingual speakers of Chinese, Russian, Korean and Spanish that subjects interviewed in the United States responded in remarkably different ways to each other in cognitive interviews due to differences in communication styles and cultural norms. These findings suggest that, to be effective, cognitive interviewing protocols and approaches may need to be tailored for different language/culture groups within or across countries. For example, recent studies have examined ways of improving the cognitive interviewing experience for Spanish-speaking respondents in the United States (Park & Goerman, 2019), and respondents outside of the United States and Europe (Kelley et al., 2015).

Time and resource limitations are also typical constraints, such that thorough, consistent pretesting is rarely part of the 3MC survey design and implementation. For example, the ESS requires that each participating country carry out a pretest to, at a minimum, check whether the translations of the questionnaire are consistent with the intended meanings, and whether CAPI (computer-assisted personal interviews/interviewing)/PAPI (paper and pencil interviews/interviewing) routings work properly. However, as de Jong and Cibelli Hibben (2018) discuss, countries vary widely in how the pretests are carried out. For example, ESS Round 7 saw substantial variation in pretesting in terms of timing (how long before fieldwork was set to begin), the number of completed interviews, the mode of interviews, whether cognitive interviewing was done, and whether or not interviews were recorded (although recommended, cognitive interviews and recording were not obligatory; see Beullens et al., 2014). As Beullens et al. (2014) report, only two countries used cognitive interviewing (Estonia and Switzerland), four countries skipped checking the translation from English into the national language, and interview recording was used only in France, Lithuania, Portugal, and Belgium.

Current best practices

While there are a number of challenges in designing and implementing pretests for 3MC surveys, de Jong and Cibelli Hibben (2018) argue that, at a minimum, some form of pretesting should be conducted in each study country. For example, an expert review of the questionnaire is relatively inexpensive and should be considered standard practice. They also recommend cognitive interviewing for all new items in the source language, at a minimum, with additional language families, major regional subgroups, and countries added as resources permit. The choice of countries/languages for cognitive interviewing could also be informed by where local experience

and expertise in cognitive interviewing is available. Results from a recent study examining how question evaluation methods compare in predicting problems suggest that the best combination of methods may be expert reviews followed by cognitive interviews (Maitland and Presser, 2016), lending support for this best practice. In addition to identifying problems, in the 3MC context, cognitive interviewing should be carried out not only to identify potential problems with questions but to investigate the constructs captured by questions (i.e., construct validity) and whether or not those same constructs are captured across various groups of respondents (Miller, 2019). In this way, as Miller (2019) notes, cognitive interviewing can help ensure validity and comparability, which is particularly essential for studies seeking to make comparative estimates (see also Braun et al., 2014). Maitland and Presser (2016) also found computer-based methods (SQP and the Question Understanding Aid (QUAID)) to be the least predictive of question problems, which suggests that these methods should be used to complement but not to replace expert reviews and cognitive interviewing. Finally, a pilot should be conducted to test all aspects of the instrument and data collection processes, ideally across all populations of interest (discussed further in Section 4.6). Additional pretesting approaches and post-study evaluation should also be carried out to the extent possible based on study goals and available resources.

The Cross-Cultural Survey Guidelines highlights various tools and standards that have been developed or adopted to maximize quality during the pretesting activities in 3MC surveys (see Caspar et al., 2016). For example, Miller et al. (2008) standardized protocol in a seven-country (eight language) cognitive interviewing study to ensure equivalence in participant recruitment, interview administration, and result documentation. Fitzgerald, Winstone, and Prestage (2014) applied the Cross-National Error Source Typology (CNEST) to categorize and analyze the results of cognitive findings to detect sources of error in the questionnaire. Rooted in quantitative methods, researchers have created "predictive systems" that linked wording features with potential issues in cross-cultural survey items, including the *Survey Quality Predictor* (Saris et al., 2011) and the *Question Appraisal System QAS-04* (Dean et al., 2007). In addition, recordings (audio, video, and computer-assisted) can be used as a monitoring tool for interviewer or respondent behavior.

Recent innovations

Cognitive interviewing

By examining the respondent's response process, cognitive interviews used in 3MC studies identify and evaluate sources of error that affect questionnaire design, cross-cultural variations in response, and translation issues. Willis (2015) provides a set of minimum standards for cognitive interviews in the area of systematic data reduction, analysis, and reporting of results. He noted that raw data from interviewer notes and respondent answers can be organized into "text summary" or coded according to respondent behaviors, themes, response patterns, and survey question features. The coding scheme can also include cross-cultural issues related to translation, sociocultural influence, and other sources. Willis proposes writing cognitive testing reports using a standardized reporting format called the Cognitive Interviewing Reporting Framework (CIRF). The CIRF applies a 10-category checklist to make clear what was done during the cognitive interviews and how conclusions were made based on procedures and results of those interviews

(Boeije & Willis, 2013). He also recommends archiving the reports in the widely accessible Q-Bank database developed by a group of U.S. Federal interagency researchers.[25]

In October 2016, the U.S. Office of Management and Budget (OMB) issued seven standards for cognitive interviews conducted by or for U.S. Federal studies. This document specifically mentioned the role of cognitive interviews in examining comparability relevant to 3MC considerations, such as "accuracy of translations" and the "equivalence" and "potential bias" associated with how different groups of respondents interpret and process the survey questions. Meeting these seven standards is necessary for Federal cognitive interview studies to be considered "accurate and trustworthy." The standards are: (1) developing a methodological plan that describes the cognitive interview study design; (2) selecting a purposeful sample to address the objectives of the study; (3) designing an interview guide that includes the survey questions to be evaluated and cognitive interview probes and instructions to conduct the interview; (4) applying systematic analysis to cognitive interview data; (5) making analysis transparent; (6) documenting methods, results, and conclusions in a final report; and (7) making final reports available to the public. When applying the OMB standards to cognitive interview studies involving non-English languages in the United States, Goerman (2017) points out the complexity in meeting the standards when it comes to sample selection, respondent recruitment, interview guide, and systematic comparisons in analysis. For example, geographic diversity of the Spanish language is a reality and should be represented among the Spanish-speaking respondents for the pretesting to be informative. Diverse Spanish speakers become hard to recruit when the study also requires that they represent specific characteristics necessary to test a survey topic. In addition, researchers must decide what type of probing and probes that could be used in the interview guide (Goerman's (2017) research showed that some probes were difficult for non-English speakers) and also how to compare the interview data across respondents speaking different languages.

Recent research addresses some of these complexities in cross-cultural cognitive interviews. Park, Sha, and Willis (2016) compare cognitive interviews conducted with respondents with different levels of English proficiency. The interviews were conducted in the respondents' native, non-English languages. The authors find that differences in responses seemed to be driven by different demographic characteristics of the respondents and not necessarily by their English language proficiency. Based on these findings, the authors recommend not restricting sample selection and recruitment to non-English speaking monolinguals, which was a common practice. Goerman et al. (2018) support this recommendation but also suggest that the sample selection decision be based on who the primary users of the translation might be. In terms of improving the interview protocol, Park and Goerman (2019) show that non-English speaking respondents can have better rapport and better understanding of the probing questions when the interviewer fostered informal interactions and explained the cognitive interview task during the introductions. In addition, Miller (2019) provides a comprehensive guide to designing a comparative cognitive interview study, including ethnographic-centered interviewing techniques, interviewer training, and development of protocols. She also introduces Q-Notes, a software designed for data entry and analysis of cognitive interviews. The analyst can use Q-Notes to

[25] See https://wwwn.cdc.gov/QBANK/Home.aspx.

examine whether or not a question performs similarly across countries, languages or any other type of subgroups.

Web probing

Web probing supplements face-to-face cognitive interviews. It involves "the implementation of probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions" (Behr et al., 2017, p. 1). Web probing involves standardized probe questions programmed to appear at strategic points through a self-administered questionnaire. Recent studies have shown that web probing is highly useful in revealing equivalence issues in surveys (Behr & Braun, 2015; Braun, Behr, & Medrano, 2018; Meitinger, 2017). In the 3MC context, web probing can be used to assess the comparability of survey items and identify possible reasons for a lack of equivalence. There are many strengths to web probing – it can reach a large and diverse sample of respondents quickly because it is administered via the web, is standardized and anonymous, and does not require local interviewer presence. However, web probing can only reach online population groups, there is probe nonresponse, and insufficient probe answers from a content perspective cannot be followed up. To harness the strengths of web probing, Behr et al. (2017) indicated that researchers must carefully take into account design considerations (probe type, probe wording, probe placement, text box size, number of probes, and the sequence of probes) and invest time and effort to develop a coding scheme that addresses the research questions, codes the probe answers, and finally checks the codes.

Usability testing

According to Geisen and Romano Bergstrom (2017), usability testing complements the other commonly used pretesting methods. Usability testing of surveys involves giving realistic tasks to users (respondents or interviewers) to reveal the ease or difficulty of navigating the survey, entering answers, and finding information to complete tasks. Like cognitive interviewing, the sample of respondents is small and not representative of the population. That, however, might not necessarily be a weakness, depending on the goal of the usability test. Usability testing is most often used in web-based and electronic surveys. Sha, Hsieh, and Goerman (2018) assessed how respondents with limited English proficiency interacted with translated prototype materials for a web survey. They found that when translation and common website functionality visual cues (tabs, hyperlinks, drop-down menus, and URLs) are presented together, they help to improve limited English-speaking respondents' experiences using and accessing entry pages and informational web pages for surveys. The authors were able to examine translation and usability at the same time because they had combined usability testing and cognitive interviewing techniques.

Suggested future directions

Measuring cross-cultural validity of pretesting results is a continuing challenge. Many commonly used qualitative pretesting techniques are not designed for the results to be generalized to a broader population. Thus, the extent to which results from such studies are reproducible and can be generalized to a broader population or to other similar cultural or language populations or population subgroups is relatively unexplored. To enable the reproduction and assessment of

pretesting *results*, the first step is making the methods transparent for public inspection and use. To that end, the new U.S. OMB guidelines, the CIRF, and the qualitative framework focusing on quality such as the Total Quality Framework (Roller & Lavrakas, 2015) have advocated transparency as a tool to advance quality.

3MC surveys are best served when researchers are able to consistently conduct measurement equivalence testing among a set of complementary pretesting methods, make informed decisions on the extent to standardize or localize approaches to interviewing, probing, and analysis across countries and languages, and to collect, make sense of, and have the opportunity to implement changes based on pretesting results. New approaches need to be developed, particularly in the area of focus groups and usability testing, because little research exists that directly addresses the cross-cultural comparability of results from these pretesting methods (see Sha et al. (2020) for an exception).

Expert review is also a pretesting technique that is commonly used but under-researched in the 3MC context. Goerman et al. (2018) discussed the goal and role of expert review in the survey translation process and described how it is currently done at the U.S. Census Bureau. They pointed out that it is important to determine more specifically which procedures are applied across organizations, what are the current best practices, and what procedures have resulted in successful expert reviews. They also observed that expert reviews are often the only form of pretesting for survey translations. The ideal approach to producing a high-quality survey, according to the authors, is to have parallel development and testing of the source and translated materials.

4.6 Field implementation

Introduction and key operational and design challenges

Fieldwork implementation encompasses a wide range of processes, including development of comprehensive interviewer tools, interviewer recruitment, training, monitoring, interview verification, and identifying and reducing nonresponse bias. Field implementation to collect comparable data in the context of 3MC surveys is one of the most challenging endeavors in survey research, particularly in face-to-face surveys, which is most often used in 3MC settings.

Factors such as geographical scope, variation in seasons and climate, timing of holidays, political regimes and election schedules, and other factors can affect timing of data collection and subsequent quality of comparisons of countries. Additionally, researchers designing surveys must anticipate but can never fully prepare for unfortunate events, including extreme weather, natural disasters, threats to interviewers, and other exogenous shocks that may require delays and/or adaptations (Pennell et al., 2014). 3MC survey projects often include countries with varying research traditions and, at times, limited exposure to state-of-the-art survey methodology current best practices. Such differences can affect every stage of the survey life cycle, including interviewer recruitment, training, assignment, and remuneration, all with a direct effect on data quality. Social desirability bias can also differ across contexts and differentially affect quality through a variety of mechanisms, including characteristics of the interview setting and of the interviewers themselves. For example, third party presence in the interview setting varies greatly,

with estimates ranging between 17% and 82% of interviews achieving limited privacy outside of the U.S. and Western Europe (Casterline & Chidambaram, 1984; Mneimneh, 2012). There is evidence that such presence affects reporting of sensitive information (Aquilino, 1993; Aquilino, Wright, & Supple, 2000; Moskowitz, 2004) and reporting can vary depending on the relationship between the respondent and the other party (Mneimneh, de Jong, & Altwaijri, 2020). Managing these varying considerations across a range of country conditions and teams of varying capabilities and practices makes data collection a major challenge for 3MC survey projects, and necessary country-specific adaptations can result in increased comparison error even though other sources of error within an individual country may be decreased by modifications to the processes.

Careful attention to detail is required at every step of the process to yield high-quality survey data. This section provides an overview of key operational and design challenges affecting data quality in 3MC survey fieldwork implementation and monitoring, industry current best practices and recent innovations including strategies for addressing both unintentional and intentional deviations from survey specifications, and future directions for improving quality in the data collection stage in 3MC surveys.

Most if not all 3MC projects operate in contexts where local organizations face a variety of challenges and have varying levels of skill in conducting representative surveys following industry best practices. Some organizations may be adept at using the latest technologies whereas others may have limited experience outside of traditional paper and pencil administration. Differences in data collection mode and related institutional capacities and local resources can have a direct effect on feasibility of various quality control processes. There is no single model that can address the challenges faced by a specific 3MC survey. Some may be able to use CAPI in all countries whereas others may be limited to PAPI, resulting in potential differences in survey quality. For example, without a data collection tool that integrates questionnaire administration with a sample management system, collection of an audit trail and related interviewer monitoring activities are limited. Further challenges may stem from the number of languages covered by the project, number of time zones across which it operates, and financial resources available, among other factors. Variation in adherence to fieldwork implementation and monitoring can lead to nonresponse error, measurement error, coverage error, and sampling error, all of which can threaten the validity of the final survey estimates, with effects magnified due to comparison error (Pennell et al., 2017; Smith, 2011, 2019a).

Current best practices

Interviewer recruitment, assignment, and training

Interviewers are responsible for the implementation of the survey design and are integral to the data collection process in interviewer-administered modes, which are most common in 3MC surveys. They are often required to perform multiple tasks with a high level of accuracy. Important criteria to consider when recruiting interviewers include an interviewer's previous experience, education and literacy, computer skills, language proficiency where applicable, and navigation skills (Jäckle et al., 2013; Lipps, 2007; Pickery & Loosveldt, 2000, 2002; Stoop et al., 2016; Vassallo et al., 2015; see also Alcser et al., 2016, for detailed discussion of interviewer

recruitment and training, including sample training guides and agenda). In the 3MC context, hiring criteria should be proposed by the central coordinating center, with country-level adaptations developed as necessary and in collaboration between the coordinating center and participating countries.

In the field, interviewer behaviors can contribute to sampling error, nonresponse error, and processing error. Interviewers can also contribute to measurement error by influencing responses through their personal attributes and their behaviors, resulting in what is often referred to as "interviewer variance," which can affect data quality differentially across countries (Beullens & Loosveldt, 2014, 2016; Blom, de Leeuw, & Hox, 2011; Groves et al., 2009; Japec, 2005; de Jong et al., 2017; Mneimneh et al., 2020). While the cluster design of most area probability sample surveys confounds the sampling and non-sampling (i.e., interviewer) variances when only one interviewer is assigned to a specific cluster, elimination of such confounding is possible if respondents are randomly assigned to interviewers. However, in practice such a fully interpenetrated design is nearly always cost prohibitive, particularly in the 3MC context. More feasible is the use of interviewing teams with at least two interviewers assigned to each primary sampling unit (PSU), which permits the estimation of measurement error introduced by the interviewer. Known as a "partially-interpenetrated design," this approach facilitates multi-level modelling in data analysis to estimate interviewer and design effects simultaneously (O'Muircheartaigh & Campanelli, 1998). Surveys in a 3MC context are also susceptible to differences in the cultural environment, existing infrastructure, and resources available (Smith, 2007). Especially problematic is the fact that interviewer variance can only be estimated via special designs and is not reflected in the regular margins of error. Thus, if this error component is not eliminated or estimated and accounted for, we will get reduced effective sample sizes and overstated confidence levels.

Ensuring training is comparable and thorough and that the requirements and expectations of the field team are clearly understood is central to not only improving data quality, but also decreasing the risk for fabrication, defined as an "intentional deviation from the stated guidelines, instructions, or sampling procedures by any member of the survey project," with fabrication of the entire substantive instrument, specific sections, or individual items possible, as well as of sample management data such as contact attempts (Robbins, 2019, p. 771). In large 3MC studies, training of all interviewers simultaneously is not feasible. A model frequently used is the "train-the-trainer" (TTT) model. Here, training is generally done in one common language. Each country or cultural group sends one or more individuals who can understand and work in the language of the trainers, to the central training. These representatives return to their own country or cultural group, adapt and translate the training materials as needed, and train their interviewers. This model allows for tailoring at the country or cultural group level, such as procedures needed for gender matching of interviewers and respondents. Separate trainings are often offered for supervisors and interviewers in this model since their roles are quite distinct. A modification of this training model is training by region or shared language if a training in one *lingua franca* is not feasible. Of course, such a model takes additional time and can result in a certain loss or distortion of information as it is passed along (Alcser et al., 2016). No matter the model used, training should emphasize that team members should seek help from supervisors, the country team, or the project leaders in case of problems or unexpected events. Additionally, it should be clear to all members of the survey team that it is imperative that everyone on the field

team must adhere to the stated requirements for fieldwork implementation. A recent overview of interviewer training in multinational programs is provided by Ackerman-Piek et al. (2020).

Pilot testing

Pilot testing plays an essential role in identifying and potentially reducing nonresponse, measurement, and processing error, all of which affect statistical estimates at the population level and thus endangers comparability in 3MC surveys. In contrast with pretesting, which focuses specifically on the questionnaire, pilot testing includes activities designed to evaluate both a survey instrument's capacity to collect the desired data and the overall adequacy of the field procedures. In the 3MC context, such testing in participating countries is crucial, although rarely done. Minimum standards should be defined, ideally in the contracting/tendering process (Orlowski et al. 2016). Success metrics should be based on precise terms and conditions of the agreed-upon design, including number of attempted/completed interviews and any quotas of specific subpopulations (e.g., five females and five males within each of three different age categories), proportion of interviewers expected to participate, regional areas, mode, and the testing time frame.

An additional benefit of conducting a pilot test is that it allows for project leaders or local teams to evaluate the performance of interviewers before actual fieldwork begins. Having all interviews participate in the pilot test is therefore an advantage. Interviewers who are not clear on the process can be retrained or, potentially, removed from the project. Some projects intentionally train more interviewers than is required to complete the survey, with the intent that those who do not meet the necessary data quality requirements can be dropped from the project before fieldwork begins, thereby minimizing potential delays in fieldwork in those countries.

Quality control

Monitoring of ongoing data collection implementation is essential to ensuring data quality across all 3MC countries and reducing various sources of error, including measurement error, nonresponse bias and data fabrication. The use of paradata to monitor survey outcomes throughout fieldwork to assess compliance and strategies to optimize contact and cooperation is crucial in reducing the total survey error. With the increasing integration of complex technology in surveys, paradata have become widely available to researchers, providing additional tools to evaluate and reduce survey error sources across participating countries (Kreuter, 2013).

Paradata is particularly useful to monitor nonresponse bias. Measuring and addressing nonresponse bias across a diversity of contexts during field implementation is challenging. Nonresponse bias is a product of the nonresponse rate and the difference between respondents and nonrespondents. The nonresponse rate refers to the proportion of eligible sample units who fail to complete an interview, while the latter refers to the differences between respondents and nonrespondents on the measures of interest. If there is no such difference between respondents and nonrespondents, then there is no nonresponse bias, regardless of the magnitude of the response rate. However, if nonrespondents differ from respondents, the lower the response rate, the higher the bias is likely to be (Groves, 2006; Groves and Peytcheva, 2008). While Groves (2006) and Groves and Peytcheva (2008) found no connection between the response rate and

nonresponse bias, Brick and Tourangeau (2017) have shown that the response rate alone can be a poor predictor of nonresponse bias, and a meta-analysis suggests that nonresponse bias in demographic estimates is not predictive of nonresponse bias in substantive estimates (Peytcheva & Groves, 2008). Unfortunately, it is difficult to assess nonresponse bias because little data is typically available about nonrespondents. One might say, though, that the risk for nonresponse bias increases with increasing nonresponse rates.

Response rates are declining in many high- and middle- income countries and researchers are increasingly concerned about the potential impact of nonresponse bias and the increasing costs associated with maintaining response rate requirements (Beullens et al., 2018; Brick & Williams, 2013; Groves, 2011; Kreuter, 2013; de Leeuw et al., 2019; Peytchev, 2013; Wagner & Stoop, 2019). Even in such as the ESS, where there is rigorous standardization in interviewing processes, response rates may differ substantially between countries (Beullens et al., 2018). Published response rates may also hide substantial differences in fieldwork operations. In the EU-LFS, for instance, the proxy rate (i.e., the percentage of responses provided by someone else in a household) shows a very large variation across countries, making response rates incomparable, and potentially making assessments of the comparability of measurements across countries even more difficult (Wagner & Stoop, 2019).

Paradata can be used from a sample management perspective as well to inform responsive designs by focusing on reducing nonresponse bias and cost. Using responsive designs, researchers continually monitor selected paradata and survey data to inform design interventions in real-time based on the error-cost tradeoff goal. Sequential phases, employing different design protocols, attempt to bring in a different set of sample members to the respondent pool. Groves and Heeringa (2006) list the following steps for employing responsive design to minimize nonresponse bias and cost:

1. Pre-identify a set of design features that may affect nonresponse error and cost, using evidence from previous waves or similar studies.
2. Identify a set of indicators of the cost and nonresponse error properties of those features and monitor those indicators in the initial phases of data collection.
3. Alter the design features of the survey in subsequent waves based on pre-identified cost-error trade-off decision rules.
4. Combine data from different design phases into a single estimator.

This strategy assumes real time monitoring of data collection for such interventions to be employed. Even when such data are being collected, however, few if any 3MC surveys are using responsive designs in the ways that Groves and Heeringa (2006) outline. This is largely due to the considerable resources needed to set up and monitor such indicators across many populations or countries. However, 3MC surveys that are repeated over time have an opportunity to use past wave data to inform contact and other strategies for future waves or to improve post survey adjustments.

The consistent trend with increasing nonresponse in middle- and high- income countries is concerning and will undoubtedly result in modified data collection strategies where multiple data sources and mixed-mode designs will be used. Currently survey organizations can utilize design

features such as respondent incentives to increase response rates, various post-data collection adjustment procedures or both (Groves & Couper 1998; Groves et al., 2002; Stoop et al., 2010). In the long run, however, other strategies must be added to address nonresponse error.

Paradata can also be used to monitor and evaluate interviewer adherence to protocols during data collection (Hyder et al., 2017; Kirgis & Lepkowski, 2013; Mneimneh et al., 2019) and to study interviewer effects and the interview context (e.g., third-party presence during the interview) in the analysis phase of a project (Benstead, 2014; Benstead & Malouche, 2015; Heeb & Gmel, 2001; Johnson & Parsons, 1994; Mneimneh et al., 2020). These data can be used to ensure that interviewers are working in the correct location and adhering to the contact and selection protocols. As mentioned above, depending on the sample design, response rates can be manipulated by not fully documenting all contact attempts, e.g., only recording successful attempts which will greatly overestimate the response rate.

Where real time monitoring is not feasible, paradata can be used to inform strategies in subsequent waves. For example, paradata and statistical algorithms can be used to optimize calling strategies (e.g., finding the best time to call or determining how many call attempts to make). ESS uses information extracted from call records, to provide feedback to fieldwork organizations for the next round of data collection and to analyze nonresponse. The contact forms allow for calculation of response rates and compare field efforts across countries. As mentioned by Stoop et al. (2010), using auxiliary data from ESS, optimal visiting time can be predicated, and respondents can be classified according to field efforts in an attempt to minimize nonresponse bias. However, real-time intervention remains a challenge given the considerable resources involved in such an effort.

Exploration of potential nonresponse bias can also be facilitated through systematic collection of paradata for both respondent and non-respondent households as well as characteristics of neighborhoods that can be used for nonresponse prevention strategies during data collection as well as nonresponse adjustment (Blom, Lynn, & Jäckle, 2008). Such data might include call records which detail call and contact attempts and outcomes, interviewer observations of doorstep interactions and/or neighborhood characteristics, and auxiliary data from external sources (Kreuter & Olson, 2013; Lepkowski et al., 2013; West, 2013). Indeed, analyses using data from the ESS have provided evidence that paradata on neighborhood characteristics may allow for correction of nonresponse bias in survey estimates (Stoop et al., 2010), indicating an increased importance in identifying and recording additional data from nonrespondents thought to be correlated with key variables of interest. That said, we note the caveat that observational data itself is vulnerable to error, and interviewer training in each participating site should include a section focused on the collection of these data, with a component where interviewers practice their skills and inter-rater reliability is assessed (Campanelli, Sturgis, & Purdon, 1997; Sinibaldi et al., 2013; West & Kreuter, 2013). However, such additional (observational) data collection is relatively low cost and has the potential for significant impact for reducing nonresponse among specific subgroups if information can be obtained about nonrespondents and used to target them to minimize nonresponse bias.

Other fieldwork implementation processes

There are additional mechanisms as well that are important to consider when implementing fieldwork processes to maximize data quality with important considerations in the 3MC context, including respondent/interviewer interactions, interviewer remuneration, respondent incentives and local review of the overall study design and field procedures for adherence to local ethical standards and laws.

First, an individual's initial contact with an interviewer can determine whether s/he becomes a respondent or a nonrespondent. Whether this initial contact takes place in person or by telephone, however, can impact the outcome of the interaction, and investigations of mode effects demonstrate that initial face-to-face contact results in greater contacts, and subsequent cooperation, than initial telephone contact (Holbrook et al., 2003; Hox & de Leeuw, 1994). Therefore, initial mode of contact should be standardized where possible in 3MC surveys so it does not induce differential nonresponse.

Interviewer remuneration can be used as a tool to incentivize country partners to increase data quality. Often, country-specific data collection organizations set daily quotas for members of their field teams. However, quotas can incentivize hurried, sloppy work or even fabrication of interviews, and differences in quotas can result in comparison error. It is more advantageous to pay an hourly rate, rewarding interviewers and supervisors for quality over quantity (Lavrakas, 1992; Pennell et al., 2010; Stoop et al., 2016; Sudman, 1966).

Similarly, payments should be structured so that positive behavior is rewarded at all levels of a country partner's organizational structure. Placing strict measures in contracts about data quality or fabrication that include monetary penalties can change the behavior of local organizations. These processes can have important implications for quality of comparative data, particularly if permitted to differ across participating countries, and it is critical that country-specific remuneration procedures are documented. However, it is also important to provide the necessary support to country leaders as well, particularly as suggestions to transition to an hourly rate are often met with resistance. Moving beyond a contractual relationship to a partnership where 3MC project leaders share techniques and best practices with organizations in each country, including on projects that may be unrelated to the work of the 3MC project, can yield increased trust and a more open relationship with local teams.

Respondent incentives can also play a contributing role to both increasing and decreasing comparison error. Nonresponse can be reduced by offering respondents an incentive for participating in a survey (Singer, 2002), and, therefore, contribute to differential response error when incentives are not used in a comparable fashion. In the 3MC context, incentives are likely to vary across participating countries based on local resources, customs, and ethical regulations (Kessler & Üstün, 2008; Wagner & Stoop, 2019), and effects may vary as well (van den Brakel et al., 2006). If an incentive is used, the amount and type, time of implementation, and any special strategy, such as increasing the amount of the incentive in the final weeks of the study, should be thoroughly documented, ideally as variables in the case-level file.

Ethical considerations may also present challenges that may be addressed at the study design stage but are operationalized in fieldwork implementation. Countries vary widely in official permissions and requirements pertaining to data collection and access as well as in regulations pertaining to ethical review and informed consent. For example, in the E.U., a new General Data Protection Regulation (GDPR) has been implemented with the intent to protect personal data in the context of linking auxiliary data to survey data (see Kolsrud, Rød, & Segadal, 2019 for a discussion). Should this regulation prove to be an efficient instrument for upholding the privacy of citizens and consumers, other countries may start adopting similar regulations. Already, many international companies with a presence in the EU have adopted GDPR compliant policies not only for their EU branches, but across all of their countries of operation. In 3MC surveys, ethical norms can also differ widely. As de Jong discusses (2019), researchers may encounter situations that require careful consideration and possible design trade-offs in order to comply with ethical principles, minimize sources of survey error, while maintaining comparability across countries or cultures. As an example, de Jong (2019) notes that maintaining sensitivity to cultural differences by having other family members present during an interview may conflict with ethical obligations to protect confidentiality and to minimize error in respondent reporting.

Recent innovations

Fieldwork process innovations

Technological changes have significantly changed the menu of methods available for overseeing the process of fieldwork, with important consequences for 3MC surveys. The availability of cost-effective devices and user-friendly software, which can facilitate comparable quality control processes across study countries, is permitting a veritable revolution in approaches to quality control in face-to-face 3MC surveys (Seligson & Moreno Morales, 2018). Particularly important in low-resource settings, such technologies deliver cost-savings in the form of efficiencies gained in fieldwork time and in the elimination of data entry costs, as well as reduction or elimination of back-checks (see Blom, 2016). The timing features of electronic devices (e-devices, e.g., smart phones, tablets, and laptop computers) can be used to audit for low quality interviews and potential fabrication (interviews that take place at unrealistic times or have improbable durations) (Seligson & Moreno Morales, 2018). Sophisticated use of available software for electronic data capture in the field can be used to build in subroutines that automatically flag suspicious interviews. E-devices enhance the researcher's capacity to capture survey metadata and identify cases with potential issues with location, contact attempts, and timing. The ability to monitor these features results in increased and systematic quality control across study countries (Montalvo, Seligson, & Zechmeister, 2019). For example, LAPOP's AmericasBarometer project programs in a "Contact Attempts" survey module prior to the interview, so that the software captures all contact attempts in such a way that the data can be verified via audits of the "electronic crumbs" dropped by interviewers on their routes and via timing data. This also facilitates information transfer of contact attempts into databases for the calculation of response rates.

The capacity of e-devices to capture additional visual and audio data can also assist in quality control. Researchers can program software to capture a front-facing picture, thus taking an image of the interviewer while the interview is in progress, to confirm that the person conducting the

interview is indeed the interviewer assigned to the case. This approach assures the researcher that devices are not passed on to individuals "subcontracted" by enterprising field teams, which may otherwise happen in some studies. The ability to capture images can also be of value when the project requires that the field team returns to the specific location, either during the field period or during the subsequent wave of a panel survey. Sound capture options on e-devices permit researchers to record audio files, with the respondent's permission, during the course of an interview for quality control and verification, although in a 3MC setting, permission for audio recording may vary across countries, resulting in varying degrees of quality control and ability to monitor and assess quality in a comparable fashion.

In many low- and middle-income countries where paper-based surveys have historically been used, interviewers are organized into small teams for fieldwork, with each team headed by one supervisor who travels with the team in the field and supervises the administration of a subset of each interviewer's assignments for quality control purposes. With the introduction of e-devices in many of these countries, another added benefit may come from a potentially changing dynamic in the role of the supervisor. Instead of serving as the project "police" seeking to detect errors by interviewers, the supervisor can spend additional time working with interviewers to improve their interviewing skills, since primary quality control falls to the central research team. Anecdotal evidence suggests that this shift in responsibilities may promote greater trust and cooperation within field teams, making interviewers more likely to seek advice from a supervisor about how to handle a problem rather than to fear sanctioning if a problem arises. However, because the introduction of technology into low- and middle-income countries is relatively recent, there is little empirical evidence about the degree to which in-field supervision is beneficial compared to remote supervision, especially as many organizations spend a significant amount of resources on in-field supervision.

The adoption of e-devices and techniques that permit the researcher to capture and utilize location, image, timing, and sound data from the field combine to bridge the gap between the researcher's office and the enumeration team's efforts in the field, which is particularly important in a 3MC survey with numerous study sites and a central coordinating center. Unlike paper-based survey instruments, the complex paradata collected by e-devices are difficult (but not always impossible) for field teams to manipulate. The resulting data, which can be uploaded daily to a central or cloud-based server, can be used to centrally monitor minor and major violations of the survey protocol, and corrections can be made while the survey is still in progress. Some 3MC projects have developed and/or deployed commercially available software that facilitate a system of automatic flags to identify suspicious interviews, allowing for greater scrutiny by the research team on a daily basis. Others use these data to develop a plan for callbacks (back checks), targeting the interviews with the greatest likelihood of anomalies. In the 3MC context, such a systematic monitoring system can indicate unusual patterns in one or more countries, facilitating timely intervention. Specific quality control analyses to detect such patterns are discussed in the following sub-section regarding fabrication control.

Yet, e-devices do not represent a silver bullet that alone is sufficient to prevent deviations from fieldwork protocols. Interviewers and supervisors also have access to technological advances, which can be used to defeat built in features of computer assisted interviewing, although this certainly can vary across study countries. For example, applications have been developed that

allow GPS locations recorded by devices to be changed, and available software varies in its ability to detect such unauthorized modifications. While technological advances offer benefits to those seeking to detect potential deviations, they also offer new opportunities to potential fabricators to avoid detection.

A potential complication with focusing efforts on preventing deviations through the use of e-devices is that this technology is not available in all contexts. Security concerns may prevent interviewers from carrying e-devices in some areas, and restrictions from relevant authorities may force the field team to use PAPI in some countries. Additionally, acquisition of GPS coordinates of interviews and audio recordings may require explicit consent depending on country-specific privacy protection laws, which can systematically affect cooperation and nonresponse bias, and more generally contribute to comparison error. Secure data storage and transfer options (e.g., through encryption) will also vary depending on the device and connectivity.

Finally, the introduction of e-devices may have unintended and differential consequences on respondent behavior. One recent study investigated whether responses collected in Wave 1 with PAPI changed when some individuals were interviewed in Wave 2 by interviewers using tablet computers. Consistent with the wealth effect hypothesis, more than half of the poorest respondents reported a higher income in the second wave when interviewers used tablets (see Bush & Prather, n.d.). While these analyses were based on a single-country survey, the results highlight the opportunity for significant differences in response patterns across modes in the 3MC context as well.

In 3MC surveys or in specific countries where use of an e-device is not an option, 3MC projects must take additional steps to ensure high data quality and prevent fabrication when using PAPI. The ability to conduct real-time checks is more limited, but 3MC projects can implement a number of safeguards to collect reliable data within such contexts. For example, the coordinating center may request partial data sets over the course of fieldwork from these countries, with delivery of data from the first respondents transmitted as quickly as possible to maximize the opportunity to address any issues. Although not received in real-time, it may still be possible to address potential issues before the completion of the survey, allowing for modifications or retraining of interviewers while data collection is still underway. Additionally, if some interviews do not meet data quality standards, respondents can be reinterviewed before the conclusion of fieldwork.

There have been recent innovations in software for project management as well. For example, the EVS implemented use of a portal called SmaP, designed by the Consortium of European Social Science Data Archives (CESSDA), for project management, data sharing, and processing. With 3MC surveys often having complex procedures, a large number of documents, and multiple teams, such innovations are critical in facilitating communication and document sharing.

Fabrication control

The multi-site nature of 3MC research contributes both to increased avenues for fabrication and to challenges in identification of malfeasance. Checking potential fabrication of data can take

place at various points within the course of fieldwork. If using CAPI, it may be possible to upload data to a central or cloud-based server automatically and conduct quality control analyses in real-time. If using PAPI, such checks may occur at the end of fieldwork or on a partial data set delivered during the course of fieldwork. In either case, it is critical for 3MC project leaders to determine which procedures or combination thereof to use to evaluate the extent of this problem. A number of statistical tests and checks are available and should be considered to control this error source in each study country, including:

- Comparing results to Benford's Law (Benford, 1938; Schäfer et al., 2004);
- Duplicate and near duplicate analysis (Kuriakose & Robbins, 2016);
- Principal component analysis (PCA) (Blasius & Thiessen, 2015);
- Multiple correspondence analysis (MCA) (Blasius & Thiessen, 2015);
- Unusual patterns in the data (Inciardi, 1981; Murphy et al., 2004; Turner et al., 2002);
- Rare response combinations (Murphy et al., 2004; Porras & English, 2004);
- Undifferentiated response patterns (Blasius & Thiessen, 2015; Inciardi, 1981; Schäfer et al., 2004);
- Short paths through the survey (Bredl, Winker, & Kötschau, 2008; Finn & Ranchhod, 2013; Hood & Bushery, 1997);
- Missing data or incomplete interviews (Bredl et al., 2008; Murphy et al., 2004; Turner et al., 2002);
- Interview duration (Bushery et al., 1999; Kresja, Davis, & Hill, 1999; Li et al., 2011);
- Duration between interviews (Robbins, 2019);
- Increasing number of interviews close to the deadline (Bushery et al., 1999; Li et al., 2011);
- Unusual times of the day for interviews (Kresja et al., 1999);
- Surge of interviews (Bushery et al., 1999; Kresja et al., 1999; Li et al., 2011);
- Missing respondent phone numbers (Turner et al., 2002; Bredl et al., 2008; Murphy et al., 2004); and
- Unusual patterns of interviewer variance (Landrock, 2017).

Thus, there are many methods and metrics available to discover potential fabrication, and it is important to find its root causes. The sheer number of methods indicate that this is a big problem for many survey organizations and ultimately for our industry. Cause for possible interviewer-initiated fabrication include difficult response rate requirements, insufficient training, low payment, payment model, the burdensome nature of the work, in addition to factors such as incomplete internalization of the project values, lack of team integration or personal motivating factors that result in such non-adherence to the study protocols. Experience shows, however, that sometimes fabrication is instigated or even performed by the data collection organizations themselves. Such causes include risk that a country might be excluded from the survey for not living up to expectations, either because some specifications are not understood or because the demands on the organization are overwhelming. It is important for the survey sponsor and the central coordination center to understand and acknowledge these root causes when determining the study protocol, particularly if mandated conditions increase the likelihood of suboptimal data collection, including fabrication.

There has been more attention in recent years to the importance of a tiered system for detection of observations deviating from expected patterns and subsequent scrutiny of said observations (see Blasius & Thiessen, 2021; Smith, 2019c; Stoop et al., 2018). Not all observations that fail one or more of these tests will be of low quality or indicate unintentional or intentional deviation from the fieldwork protocol. Valid explanations can exist in some cases whereby legitimate errors occurred or unusual patterns can be justified. 3MC project leaders should not automatically reject such observations, but instead seek to understand the process that created such errors and then determine the steps that should be taken to address any issues. Indeed, cases that are ultimately rejected can also be delivered as part of a deleted-case file to be analyzed as part of a post-project assessment of fieldwork protocols. Different combinations of these analyses have been used to detect deviations in 3MC surveys with mixed results (Bergmann & Schuler, 2019; Malter, 2017; Stoop et al., 2018), illuminating a field of future research as noted below.

Suggested future directions

While innovations are critical for data quality improvement, challenges to the implementation of these approaches remain for both field staff and researchers, with significant opportunity for improvement. With respect to the former, field teams in at least some 3MC study countries may be reluctant to adopt new technologies or to change their existing methods. There are obvious start-up costs involved, including the purchase of the hardware and training at all levels of the organization. While e-devices lead to efficiencies in fieldwork and, thus, cost-savings over time, data collection organizations in specific countries may insist that researchers pay the costs of hardware and additional training for their fieldwork staff in at least the early stages of adoption. The survey research team and/or central coordinating center must take deliberate steps toward working with survey organizations regarding these concerns and provide the support required to hone supervisors' and interviewers' skills with the new devices, software, and protocols for electronic data capture in the field.

Another area for quality improvement relates to privacy concerns for the respondent. For example, respondents may be reluctant to provide consent for audio recording, and in some countries, there may be legal obstacles as well. Such variation can lead to differences in nonresponse, nonresponse bias, and/or, more generally, differences in quality control procedures across study countries. Additionally, having precise GPS coordinates for interviews may make it possible for the respondent to be identified if sufficient safeguards are not taken by the local team and the research team to prevent such possibilities. Again, the survey research team and/or central coordinating center must look for strategies to address these concerns.

Furthermore, field teams might rightly be concerned about security when carrying e-devices and researchers should consider whether, in certain instances, the sampled area is too "hot" for data collection. There are such areas in virtually all countries. Consideration of interviewer safety when carrying e-devices is therefore universal. Of course, even minimal increases in risk need to be seriously considered by the researcher, who should work with the survey team to minimize risks (e.g., the use of camouflaging cases to shield e-devices from public view). Moreover, in cases where the use of an e-device is not possible or may pose an undue risk to the interviewer, it

is important that local partners and research teams maintain the necessary skills to perform oversight for and data quality analysis using a paper-based instrument.

For the researcher and/or central coordinating center, the adoption of a CAPI system that permits extensive quality control requires investment in acquiring the knowledge and skills to develop strong protocols for development and implementation, revision of training materials, development of software and other quality control analyses that exploit the full capacity of e-devices, and providing technical support to teams in the field. Researchers deploying e-devices for fieldwork also face the challenge of staying one step ahead of creative minds who seek ways to circumvent the tight control offered by e-devices.

Devices need to be tested and often de-bugged prior to each study and in each country, which becomes increasingly complicated as more sophisticated techniques and methods for designing survey instruments become common. Additionally, researchers need to prepare to lend support throughout the duration of fieldwork to teams that encounter technical problems along the way. In the 3MC context, the need for such support is likely to vary greatly across countries, depending on resources available in the country. For the researcher, as well as the survey team, there are large start-up costs – in time and money – required for the effective implementation of e-devices for fieldwork implementation and quality control. Indeed, currently there are few low-cost data collection platforms that integrate a sample management system with a data collection tool, which also permit collection of complex paradata. Investment in development of such resources is a necessary next step.

Although challenges remain in implementation, these recent innovations provide a number of approaches and tools to help local organizations to collect the highest quality data possible under given country conditions. However, all of these processes require project leaders to devote significant time and resources across the survey lifecycle. Building these costs into the project is imperative from the onset and will require critical effort in capacity building in those countries within a 3MC project that have a more limited survey research tradition so that data of comparable quality can be collected.

Improving data quality also requires 3MC projects to seek to shift the narrative about data quality, and especially about intentional deviations from protocol and outright fabrication in the field of survey research. Claims of fabrication have been equated with an attempt to delegitimize a particular survey, rather than recognizing that such intentional deviation can be a common, albeit regrettable, form of survey error that can affect any 3MC project. In effect, fabrication represents yet another source of survey error that can bias estimates at different stages. The main difference is that it represents *intentional* introduction of error as opposed to the unintentional forms of nonsampling error detailed in a TSE framework.

The default assumption that data are correct unless definitively proven otherwise is counterproductive for improving data quality in survey research. Rather, openly allowing for the possibility of fabrication or other data quality issues within the survey lifecycle offers the opportunity to change the narrative about the 3MC field. Critically, it is important for researchers to determine the process that is most likely to have produced the patterns observed in the data, including fabrication.

As discussed above, the use of paradata for strategies to reduce nonresponse bias, such as responsive design, is critically important. However, even if responsive designs are employed, it is highly unlikely that the same set of interventions will be used and/or will suffice across all participating countries. Here, we can expect each targeted population or country to respond differently to interventions given the heterogeneity across sample designs, mode of contact, at-home patterns, survey climate, survey topic and sponsoring agency, interviewer experience and training, and the proportion of inaccessible areas and buildings, among many other contextual factors and population characteristics. The data available for post-survey adjustment will also vary considerably. Given this, reducing survey error must start at the population or country level. This is not to say that important insights on reducing one type of error in one context cannot be useful in another (for example, see Wagner and Stoop (2019) for a discussion of nonresponse bias reduction).

The complexity of 3MC surveys has resulted in uneven attention to aspects of the survey lifecycle more generally, and fieldwork implementation specifically. For example, while there is empirical evidence that interviewers can differentially affect data quality, and that this effect can differ across respondents, countries, and cultural contexts (Benstead, 2014; Benstead and Malouche, 2015; de Jong et al., 2017; Heeb & Gmel, 2001; Johnson and Parsons, 1994; Mneimneh et al., 2015) there has been scant research about how best to measure interviewer effects and use the data both to guide changes in interviewer training methods as well as in analyses. Such research and implementation of findings in future data collections is imperative for achieving high quality data.

Finally, communication and sharing of knowledge across projects is crucial in the quest for high-quality data. As technology progresses, the methods used both to conduct and detect deviations also change. Detailing new methods as they develop allows other projects to implement necessary safeguards and limits the degree to which newly developed methods of fabrication are likely to be successful. Ultimately, changing the cost-benefit calculation of potential deviation remains one of the efficient means to increase data quality in the fieldwork implementation phase of survey research.

4.7 Documentation in 3MC surveys

Introduction and key operational challenges

Survey data documentation is a complex concept that covers both the process whereby survey production is recorded, and the various outputs produced along the way, such as contextual descriptions, culture-specific information relevant for fieldwork, sample properties, among many others (for this definition of documentation, see Tomescu-Dubrow and Granda (2019)). The purpose of documentation is straightforward: it enables knowledge about the survey to accumulate, and to flow *within* the data production network, and *from* the network *to* data consumers (Kallas & Linardis, 2010; Niu & Hedstrom, 2008; Ruggles, 2018; Vardigan, Granda, & Hoelter, 2016).

Thus, documentation links intrinsically to data quality, as defined in the frameworks of TSE, monitoring survey quality, and fitness for use (see Section 3). Data producers need accurate,

comprehensive and timely information about the survey process, including process quality metrics, i.e., paradata, to (a) perform ongoing quality control during all stages of survey production to reduce errors related to representation, measurement, processing, and, in 3MC surveys, comparability, (b) increase management quality, and (c) render the survey interpretable to the broader research community.

Secondary users need thorough, accurate documentation to conduct informed survey data analysis. Documentation enables researchers to evaluate a given 3MC survey along important quality dimensions that TSE and survey process quality management identify, including comparability, relevance, accuracy, timeliness, accessibility, interpretability, and coherence (for details on quality dimensions, see Hansen et al., 2016; other quality dimension vectors are possible depending on user requirements, see Lyberg, 2012). At the same time, documentation is key to data curation, which presupposes integrating data across time and sources, sustainability (including preservation of, and broad access to, old survey projects), and dissemination (Ruggles, 2018).

However, consistently high-quality documentation is not yet the norm in the 3MC context (Kołczyńska & Schoene, 2019; Mohler & Uher, 2003; Mohler, Pennell, & Hubbard, 2008; Mohler et al., 2010; Oleksyienko et al., 2019; Tofangsazi & Lavryk, 2018). This is due to a combination of reasons, including (i) operational challenges of survey documentation; (ii) unequal resources and research infrastructures within 3MC projects (e.g., among country teams), and between 3MC projects; and (iii) the lack of widely agreed upon standards and best practices for documentation.

Operational challenges occur from both understandings (process and outcome) that the concept "survey documentation" carries. First, as process, documentation is inherently dynamic. This poses difficulties for data producers, who need to follow the survey lifecycle as it unfolds, in order to chronicle the series of interlinked and often iterative stages, from initial study planning to the production of final data files, and data dissemination (see survey lifecycle diagram in Figure 2 of this report (Survey Research Center, 2016)). However, dynamic documentation is more demanding in 3MC projects. In consequence, 3MC projects rarely exhibit common standards about *when* documenting occurs (e.g., on continuous basis/post factum), despite the longstanding recommendation that information should be captured at the source (Granda & Blasczyk, 2016).

Second, as output of the recording process – the available information about a given survey – documentation raises not only substantive but also terminological challenges. For example, the literature refers to 'data about data' as 'metadata' (Bargmeyer & Gillman, 2000; Mohler et al., 2008; Ruggles, 2018; Sundgren, 1973, 1995; Vardigan et al., 2016; see also Riley, 2017). Under this definition, survey metadata and survey documentation are one and the same, and the concepts are used interchangeably (Hoelter, Pienta & Lyle, 2016; Niu & Hedstrom, 2008; Vardigan et al., 2016). Yet, some authors apply the term metadata to a specific type of documentation, namely machine-processable structured documentation (Ruggles, 2018, p. 303), while others treat them as conceptually distinct (Mohler et al., 2008, p. 404-405).

The substantive difficulties link to *what* the documentation of 3MC studies should cover, and how detailed the information should be. First, in comparative survey projects, the metadata elements (e.g., study methodology report, survey instrument, codebook and correspondingly, the variable and value labels attached to the numeric data in the computer files, paradata, etc.) that characterize any single-population, one-time survey, cumulate across surveys and, often, across waves. Second, the metadata for 3MC projects are more varied. On the one hand, 3MC studies need metadata that capture harmonization decisions at the stages of design, implementation, and data processing (*ex-ante* harmonization), or after data release (*ex-post* harmonization). On the other hand, 3MC studies require metadata that capture "culture-specific collateral information, culture-specific questionnaires, and culture-specific data collection implementation practices and protocols" for as many surveys as present in a given 3MC project (Mohler et al., 2010, p. 310; Vardigan et al., 2016; see Mohler & Uher, 2003 on the need to describe the socio-cultural contexts, including events around fieldwork, that are likely to influence respondents' answers).

Third, it is difficult to decide how much detail the survey metadata should provide when preparing documentation for future users whose identities and objectives are unknown (Niu & Hedstrom, 2008), and whose familiarity with the socio-cultural contexts the data cover will be uneven (Bethlehem et al., 2008; Lynn, Japec, & Lyberg, 2006).

Challenges pertaining to when 3MC documentation is produced, what it should cover, and in what format (e.g., machine-actionable metadata) interact with the organizational structure of comparative survey projects (see also Section 4.1). For example, the responsibility to compile the different components of the 3MC survey documentation rests with multiple actors, including but not limited to, the study's director and one or more data collection organizations. Among these actors, the resources to implement and enforce common standards, even when they exist (e.g., AAPOR's standard definition and calculation method of response rates (American Association for Public Opinion Research, 2016)) are frequently unequal.

Variability in 3MC survey documentation standards

The documentation of 3MC survey projects varies widely in content, nomenclature, format, and access (Kołczyńska, 2014; Kołczyńska & Schoene, 2019; Mohler et al., 2008; Mohler at al., 2010; Oleksyienko et al., 2019; Ruggles, 2018; Scholz & Heller, 2009; Smith, Fisher, & Heath, 2011; Tofangsazi & Lavryk, 2018; Tomescu-Dubrow, Słomczyński, & Kołczyńska 2017; Vardigan et al., 2016). We discuss below findings stemming from the Survey Data Recycling (SDR) Project.[26]

The SDR Project (see also Section 3) exemplifies how intrinsic survey documentation is to secondary users of 3MC data. A main purpose of the SDR Project is to extend *ex-post* harmonization of cross-national survey data initiated in the Harmonization Project (Słomczyński et al., 2016; Tomescu-Dubrow & Słomczyński, 2016) and the SDR database v. 1.0 (Słomczyński et al., 2017a).[27] To create common (target) variables it was essential to first understand the properties of the selected source data. Since the SDR and Harmonization Projects reprocessed information from thousands of national surveys stemming from 23 international projects,

[26] See asc.ohio-state.edu/dataharmonization, NSF SMA-1738502).
[27] See SDR Master Box, https://doi.org/10.7910/DVN/VWGF5Q, Harvard Dataverse, V1.

including WVS, ISSP, LAPOP, Afrobarometer, ESS, Eurobarometer, and so on, the projects developed methodology to systematically evaluate the documentation for all these sources, and then created variables measuring documentation quality.

Between 2014-2015, Kołczyńska and Schoene (2019) coded and analyzed the English-language technical reports, codebooks, questionnaires, and other materials, corresponding to 1,721 national surveys (81 source data files, 22 international survey projects) that contributed source data to the SDR database v.1.0. The general documentation lacked discussion of pretesting the questionnaire prior to fieldwork for 68% of the surveys, for 62% it lacked information on fieldwork control, for 49% response rates were not reported, and for 28%, information on sampling was either missing or so poor that identifying a sample type was not possible (for the latter, see Kołczyńska, 2018). On the bright side, Kołczyńska and Schoene (2019) found a clear increase in average quality of documentation over time, especially since the 1990s.

During the same period, and using the SDR database v.1.0, Oleksiyenko et al. (2019) cross-checked – for gender, age, year of birth, education levels, years of schooling, trust in parliament, and participation in demonstrations – the documentation with the data records in the datasets. They identified processing errors in 20% of the examined variables. During 1968-1989 the number of processing errors in all surveys under analysis was the lowest. From early 1990 until 2007 the number of errors grew steadily. A slight improvement started to show from 2008 to 2013.

A second round of documentation evaluation started in 2017, when data reprocessing in the SDR Project broadened to include variables from a total of 3485 national surveys stored in 215 source datasets of 23 cross-national projects (SDR database v.2.0). Tofangsazi and Lavryk (2018), who, among others, coded the English-language documentation available for these source data, reflected on the difficulties that unequal standards in documentation format pose for information search. Existing technology to create machine-actionable structured documentation, such as the Data Documentation Initiative (DDI, www.ddialliance.org/), discussed further below, has yet to be widely adopted by 3MC survey projects. Its use in 3MC projects is uneven, possibly because it calls for specialized personnel, itself a likely source of additional resource inequality among participating countries in cross-national projects.

Current best practices

Current best practices for documentation of 3MC survey projects represent the result of concerted efforts to maximize data usability for secondary analyses and data curation (Dale, Arber & Procter, 1988; Mohler et al., 2010; Ruggles, 2018; Vardigan et al., 2016), while strengthening survey process quality management (e.g., Biemer & Lyberg, 2003; Lyberg & Stukel, 2010). They include a growing trend to account for the dynamic nature of the documentation process, by placing increased emphasis on the importance of "capturing metadata at the source." Producing 3MC documentation during the course of the survey lifecycle would result in richer documentation with less burden on the data producer, since information is captured as each survey is planned, questionnaires are created, and data collection commences.

To discuss best practices for documentation content, we largely organize metadata elements around a classification that data archives employ (Gutmann et al., 2004; see also Kallas & Linardis, 2010). Archiving institutions, such as the Inter-university Consortium for Political and Social Research (ICPSR), GESIS, Roper, and the UK Data Archive, among others, play a key role in preserving and sharing social science survey data. The schema we use includes: (1) study-level metadata; (2) file-level metadata; (3) variable-level metadata; (4) administrative and structural metadata; (5) paradata; and (6) ancillary documentation. Best practices presented below build on exiting knowledge about documenting elements common to all surveys, and morph to account for the specificity of 3MC survey data (see also Hansen et al., 2016, Appendix B).

Study-level metadata (sometimes referred to as project-level metadata) describe the survey project as a whole, providing metadata for each unit or series of units (e.g., national and/or subnational surveys) that form a 3MC project. In multi-wave studies, study level metadata are compiled for each unit (e.g., country) of the project-wave, and disseminated as wave-level metadata, sometimes together with the unit-level metadata (e.g., country-specific documentation). Study-level metadata include information about the project team, funding sources, project specifics, data collector/producer, study design, target population and if applicable, the survey population, the unit(s) of observation, the sampling and sampling procedures, incentives, interview mode(s), data collection instrument/instrument versions, detailed information about pretesting and translation, interviewer training, fieldwork execution and monitoring, response rates, weighting, dates and geographic location of data collection, possible additional data sources. See Appendix 6 for a detailed list.

File-level metadata describe the properties of individual files in a data collection (Gutmann et al., 2004, p. 217). The main elements include the technical characteristics of the file itself, including size, number of variables, number of cases, and additions such as a checksum to certify the authenticity of the original data file in case it might get damaged in subsequent transfers.[28]

Producers of 3MC survey projects frequently disseminate multiple public-use datasets that are at varying levels of file integration. For example, multi-wave international survey projects produce individual country-level public-use files, wave/round-level public-use files containing all or selected individual country-level surveys that were conducted as part of that wave/round, as well as multi-wave integrated datasets. All these files require their own metadata.

Variable-level metadata describe individual variables or groups of variables. Variables are typically documented via a codebook, variable labels, and value labels corresponding to data records in the computer files. Information for each variable is detailed, including the exact question wording or exact meaning of the datum, a link between the variable and the question, information about who was actually asked the question, the exact meaning of codes (i.e., variable values), missing data codes, unweighted frequency distribution or summary statistics, imputation and editing information, cumulative scaling, details on constructed variables, details on weighted

[28] A check-sum is "a digit representing the sum of the correct digits in a piece of stored or transmitted digital data, against which later comparisons can be made to detect errors in the data" (Gutmann et al., 2004, p. 217).

variables, location in the data file, and variable groupings. Please see Appendix 6 for a detailed list.

Administrative and structural metadata are mainly the purview of archives. They provide information on how electronic data collections were produced and how they could "be migrated or emulated in an evolving technological environment" (Gutmann et al., 2004, p. 217). They include technical information on files (file formats, file linking, etc.) and an alphabetized list, index or table of contents of variables with corresponding page numbers in the codebook.

Paradata are auxiliary data collected in a survey that describe the process of survey production, including, but not limited to, data collection (Beaumont, 2005; Couper, 1998; Couper & Lyberg, 2005; Kreuter, 2017; Kreuter & Casas-Cordero, 2010; Kreuter, Couper & Lyberg, 2010; Mohler et al., 2008; Morganstein & Marker 1997; West, 2011). They can be generated at all survey stages and stored as part of the dataset containing respondents' answers, or in separate files.

Paradata need to be documented. Analysts concerned with survey design and survey quality are especially interested in paradata because they enable research on sources of error, for example measurement error, nonresponse error, or adjustment error (Kreuter, 2017). Technological advances such as computer-assisted interviewing enable automated generation of paradata, while interviewers can provide rich observational data (e.g., about respondents' demographics, neighborhood conditions, and so on). Among collected and used variables are measures of response time to given questions; variables derived from the household roster obtained from screening interviews; variables derived from interviewer-generated call records when attempting to contact possible respondents and main interview attempts; observations collected by interviewers as they observe neighborhoods, housing units, and sample persons; characteristics of the interviewers themselves possibly including some demographic variables; and geographic data on sampled areas.

Some paradata files are publicly available (e.g., the ESS contact files),[29] while other files can only be made available in secure research environments under controlled conditions. One such example is the U.S. National Survey of Family Growth (NSFG) study.[30] Secondary analysts can access its paradata file in a government research data center. However, the project's website provides a general description of the paradata and links to the "user guide" and variable file index.

Ancillary documentation includes information such as data collection instrument(s) in all languages used in survey administration, the interviewer guide with details on how interviews were administered (including probes), interviewer specifications, use of visual aids (e.g., show cards), a flow chart showing which respondents were asked which questions and how various items link to each other, a list of abbreviations and other conventions used in variable names and labels, logic for recoded variables, coding instruments (definitions and coding rules), and a list of citations to publications based on the data, by the principal investigators or others. See Appendix 6 for a detailed list.

[29] See www.europeansocialsurvey.org/download.html?file=ESS9CFe01&y=2018
[30] See www.cdc.gov/nchs/nsfg/nsfg_2011_2015_puf.htm

Other standards

Although not focused on 3MC surveys, AAPOR's Transparency Initiative promotes disclosure of research methods in publicly released data. Practicing transparency means being willing to make research methods available for public inspection in the reporting of survey findings. In AAPOR's case, the list of organizations that have pledged transparency and agree to regular audits is publicly acknowledged, lending to their credibility, although AAPOR indicates that it does not equate transparency with the quality of the methods being disclosed.[31]

The European Statistical System's Standard for Quality Reports (Eurostat 2009) stipulates transparency in the context of equity. That is, users must have equal opportunity to retrieve the data, receive the information in a nonpartisan, objective manner, and be made aware of confidentiality provisions and error corrections.

ISO 20252 (2019) is the latest version of the process standard for market, opinion, and social research. Many of the requirements involve transparency of methods and documentation of those methods. ISO 9001 (International Organization for Standardization, 2015b) is the latest version of the standard for quality management systems. Some funding organizations are requiring ISO certification to bid on projects.

Recent innovations

Collaborations within international research infrastructures, such as the Consortium of European Social Science Archives (CESSDA)[32] and The Data Documentation Initiative Alliance (DDI)[33] have resulted in the development of free, online accessible, tools to facilitate the implementation of common standards for 3MC survey documentation. This is important for reaching greater agreement on shared terminology that documentation providers use (i.e., building a common nomenclature for documentation).

CESSDA, for example, as part of its data management plan, provides documentation guidelines.[34] DDI promotes tools that allow data producers to create machine-actionable structured metadata elements that also use a "controlled vocabulary." Current versions of DDI are able to fully document all project and variable-level elements of 3MC surveys, including their comparative structures, and permit machine-actionability of all contents. The possible applications of this markup are manifold, e.g., use by search engines within the individual project

[31] AAPOR's 12 basic disclosure elements for transparency are as follows (details are available on aapor.org): 1) Study sponsor and conductor; 2) Question wording and presentation; 3) A definition of the population under study and its geographic location; 4) Dates of data collection; 5) Sampling frame(s) and its coverage of the target population; 6) Name of the sample supplier; 7) (pre-recruited panel or pool) Participant recruitment methods; 8) Sample design, including respondent selection, recruitment, or contacts, and using probability or non-probability methods; 9) Survey modes and languages; 10) Sample sizes and the estimates of sampling error (probability surveys) and only measures of precision with a description of the model used for non-probability surveys; 11) Weights calculation; and 12) Study contact.

[32] See https://www.cessda.eu/

[33] See https://ddialliance.org/

[34] See https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

or across the web to facilitate data discovery; input into data analysis systems or web interfaces to encourage initial interaction with the basic features of 3MC datasets before more comprehensive statistical analyses are planned, or the production of project documentation and reports in almost any manner desired through the use of stylesheets.

The DDI structure is particularly valuable for describing 3MC projects because it is possible to track and compare individual questions over time and over countries, display identical or similar questions used in different countries on a single screen to assess comparability, and provide links to documentation from individual countries which may be in different languages.

As noted in the Questionnaire Design section, the ESS uses the QDDT, which is based on DDI and has been developed based on the ESS's complex questionnaire design process. It is a free web-based tool that helps researchers to develop thematic questionnaire modules (concepts and questions) and also captures the development history of survey items (for details see Orten et al., 2018).

Suggested future directions

To strengthen the quality of 3MC survey documentation, and thus, the quality of the 3MC survey project as an end-product, there are steps that data producers can take to reach greater consistency in implementing common standards of the documentation process and its outcomes. At the same time, international research infrastructures can contribute technological developments, open-source software tools especially, and freely share them online to lessen the burden of documentation on data producers and to facilitate dissemination of information. Finally, users of 3MC survey data can strengthen their contribution to the conversation on data quality.

Data producers

Data producers should generate documentation from the very beginning of 3MC projects and throughout the survey process, and, if needed, update and revise it on a continuing basis. In this way, it will be less burdensome to produce knowledge that is vital for both internal survey process quality management, and for informed use of the survey data by actors external to the data production network.

Going forward, 3MC data producers should focus on two interrelated areas: (i) the study's capacity for generating comparative analyses; and (ii) strengths and limitations of the study's methodology.

(i) <u>Facilitate comparability assessments</u>

3MC survey projects collect information from more than one nation, culture, and/or region with the explicit purpose of facilitating comparative research. Users want to compare data across entities – not only when producers release merged files with respondents' answers pooled into a single dataset, but also when they release a number of separate public-use files, each containing only respondents from a single nation, culture, or region. Researchers should be able to evaluate

how feasible it is to do so, including via access to descriptions of culture-specific events and contexts around fieldwork and that might influence respondents' answers.

Access to harmonization metadata is essential to assess comparability of concepts and constructs, representation, and measurement. Data providers need to describe the decisions taken – via *input* harmonization before fieldwork begins, via *ex-ante* output harmonization during data processing, or both (see Section 2.2 and 2.3). The same requirement of careful documentation applies to datasets containing variables harmonized *ex-post*.

Bearing in mind the salience of research transparency and replicability, it is especially important that harmonization metadata detail any transformations that variables in 3MC datafiles undergo prior to public release. Transformations can be extensive, for example when data producers map individual responses to comparable questions from different surveys (e.g., rating scales of different length) into a common coding scheme applicable to all respondents. Sharing the appropriate code (syntax) for harmonization is strongly recommended.

Data providers should store harmonization metadata pertaining to the harmonization process (e.g., properties of the original (source) variables, such as semantic differences, characteristics of the original scales, etc.) as methodological indicators as well (Słomczyński & Tomescu-Dubrow, 2019). "Harmonization controls" should be available in the 3MC dataset with the harmonized variable they characterize, just as imputation "flags" are commonly included in survey datafiles to alert users to values that are not original but imputed. Documentation is also needed for these controls.

Secondary users should have the opportunity to do their own data harmonization. Thus, when integrated datasets are provided, the individual, original data files from which they were produced should be identified and accessible to the research community.

(ii) Facilitate assessments of strengths and limitations in a study's methodology

A frank account of the strengths and limitations of a study's methodology will help users make informed decisions about the scope of the analyses given 3MC datasets permit. Equally important, it will facilitate new data collection that builds on accumulated knowledge.

Metadata elements pertaining to representation and measurement play an important role for methodology assessments, while also bearing directly on comparability. The SDR Project highlighted that users, to decide whether the data of a 3MC study fit or not their research purpose, should have information on at least the following survey characteristics (SDR Team, 2019):[35]

- Target population
- Sampling design and sample characteristics
- Weights
- Response rate

[35] See asc.ohio-state.edu/dataharmonization/about/events/building-multi-source-databases-december-2019/

- Questionnaire translation (method, including translators' formal expertise)
- Instrument pretesting
- Interview mode
- Fieldwork control

We add questionnaire adaptation here, and stress that survey instruments (questionnaires) should be available both in the language they were first developed in (e.g., English), and in the languages to which they were translated. When describing the features of a survey, data producers could emphasize what information could be relevant in order to consider possible room for intentional error (self-completion surveys, for instance, do not run the risk of data fabrication by interviewers, call-backs to respondents can identify interviewer fraud, etc.).

Key metadata elements pertaining to representation and measurement should also be stored as methodological variables in the 3MC datasets they characterize (Słomczyński & Tomescu-Dubrow, 2019), similarly to paradata. First, this will allow users to quickly identify main survey properties. Second, researchers can use these indicators to empirically examine the extent to which the methodological context of the survey influences analytic outcomes, for example with regards to specific components of TSE. This approach fits into, and extends, existing and recommended practices of constructing and sharing paradata.

The metadata necessary for (i) comparability assessments and (ii) assessments of strengths and limitations of a study's methodology are complementary and sometimes overlap. Together with information about corrective actions taken, lessons learned, and recommendations for improvement and further research, they provide insight into the overall quality of the survey project.

Since, on the one hand, increasingly complex 3MC designs equate increasingly complex documentation, and on the other hand, not all users need the same degree of documentation detail, data producers should synthesize the survey metadata and summarize them into a methodological profile of the study, as well as into concise "what you absolutely have to know" briefs. While practiced by some 3MC survey projects (see Hansen et al., 2016 for discussion and examples), methodology profiles, known in the literature as quality profiles (Granda & Blasczyk, 2016), should become the norm. Harmonization control variables, measures of the study's methodological context, and paradata should be part of the methodology profile, to enable effective evaluation of within and between-project characteristics.

International research infrastructures

By the very nature of comparative survey research, creating 3MC documentation requires substantial effort on the part of data producers, and increasingly so when resources are scarce. International social science research infrastructures such as CESSDA and the DDI Alliance, in collaboration with professional organizations like AAPOR, WAPOR, ESRA, and CSDI, can contribute know-how and technological developments, especially open-source software tools, to aid the production of high-quality documentation, and ultimately, that of high-quality 3MC survey projects.

One important contribution would be to develop and share best practices on documenting harmonization processes that are intrinsic to 3MC survey production. Recommendations about describing *input* and *ex-ante* output survey data harmonization procedures are currently lacking from the guidelines that CESSDA or the DDI Alliance provide. The need to create and share best practices of documentation extends to harmonizing survey data *ex-post*, both when *ex-post* harmonization is conducted within a given 3MC study (e.g., the Luxembourg Income Study), and when it constitutes a stand-alone study, like IPUMS International and the SDR Project.[36]

Relatedly, social science infrastructures can contribute further technological developments that are open-source, user-friendly and can be disseminated freely. Software that "surveys" data producers about the survey process, and stores answers as variables, would boost a standardized implementation of best practices for 3MC documentation, harmonization included. Ultimately, it would benefit secondary data use, by providing researchers the means to quickly overview some of the data's main features. The Survey Metadata Documentation System (Mohler et al., 2008; Mohler et al., 2010) could be regarded as a steppingstone for developing such an approach.

Last but not least, social science infrastructures can spur capacity building among data producers, survey methodologists and secondary users, unequal needs and resources notwithstanding. For example, Massive Open Online Courses[37] and webinars could be developed and advertised. The European Master of Official Statistics[38]could conduct courses on 3MC surveys, to highlight that quality in official statistics in many cases downplays the complexity of comparability. International organizations could follow the example of the OECD to bring together experts to talk about methodological issues.[39] Producing short (5 to 10 minute) movies (like Technology, Entertainment, Design (TED) talks) on specific methodological and documentation issues would also broaden outreach.

Data users

Secondary users constitute the quintessential audience for the information that 3MC documentation conveys about survey production, a process they are generally not part of. Yet being external to data collection does not confine users to passivity.

First, data users are intrinsic to documentation meeting its purpose to transfer knowledge. Even the most comprehensive and accurate description of a study will be instructive only to the extent to which people read it. As 3MC projects become more complex and their methodology more transparent, they provide richer documentation. This invites users to invest the effort in finding and absorbing the information (e.g., about target population, sample design and sample characteristics, response rates, fieldwork control, translation checks, etc.) that exploration of datafiles only cannot convey.

[36] See https://international.ipums.org/international/ for more information on IPUMS International. For documentation in SDR, see the SDR Team presentations, 2019 at https://www.asc.ohio-state.edu/dataharmonization/about/events/building-multi-source-databases-december-2019/.
[37] See https://www.mooc.org/
[38] See https://ec.europa.eu/eurostat/cros/content/emos_en
[39] See www.oecd.org/skills/piaac/neweventspage.htm

Second, users can contribute more to improving implementation of documentation standards. There are various platforms of communication for users to voice their experiences with sufficiency and ease-of-use of the documentation for given 3MC projects. The scientific meetings devoted to comparative survey research that major professional organizations, such as AAPOR, WAPOR, ESRA, continuously organize are one example of such discussion forums. Publications are another.

These points build up to inviting users of 3MC surveys to become a stronger voice in the conversation on data quality, to which they are an intrinsic partner. It is users who ultimately choose, from the wealth of available datasets, which one(s) to analyze. The decision requires a careful assessment of the properties of the data in light of one's specific research needs and carries substantial weight, since all audiences – academia, policymakers, NGOs, journalists and the general public – share one key expectation: that survey-driven results, and in the end, the conclusions they inform, are robust.

5. The changing survey landscape

Section 5 discusses trends in survey research such as increasing costs, reduced respondent participation and new approaches to address these and other issues and how they might apply in a 3MC context.

There are significant changes on the near horizon or already underway in survey research methodology. 3MC research will be affected by these and therefore must keep abreast of them to assess how they can be adapted to a wide variety of survey contexts. As mentioned above, 3MC advances have stemmed from many academic origins. For example, 3MC surveys conducted by organizations rooted in different academic disciplines and with varying research traditions tend to focus on different aspects of measurement and methodology. Assessment surveys, such as PISA and PIAAC, pay extensive attention to psychometric qualities of questions and assessment instruments, official statistics tend to focus on sampling, coverage and nonresponse, health surveys emphasize validated measurement instruments, academic surveys acknowledge the importance of questionnaire testing and surveys from the market research world advertise timeliness as an important asset. For all involved, survey costs are a big issue.

Recent years have seen a sharp increase in the costs of surveys, partly caused by the need to deploy more efforts to contact people and persuade them to participate (European Commission, 2018). New developments such as the GDPR in the EU also make surveys more difficult, for instance, because it makes it more difficult to supplement sampling frames with telephone numbers and background information.

Partly because of increasing survey costs, many high-income countries have moved to web surveys. One consequence is that in several European countries only one or two organizations are able and willing to bid for a high-quality high-effort face-to-face survey such as the ESS. Experience and competence in running probability face-to-face surveys is getting scarce, and it can be expected that in the near future face-to-face surveys will become prohibitively expensive in some countries. In many others, however, it is still the only viable mode option for the foreseeable future.

This means we will need to be prepared to move to modes other than face-to-face and telephone interviewing, as thoroughly described in a recent AAPOR Task Force report (Olson et al., 2019) and to consider different types of mixed-mode surveys. Mode and mixed-mode issues have been on the ESS agenda for at least 15 years, so far resulting in the decision to remain with face-to-face interviewing to preserve comparability. To move things forward, ESS undertook the initiative to design a Cross-National Online Survey (CRONOS), a pilot probability-based panel mounted in three countries. Following this, the ESS was awarded the SUSTAIN 2 Horizon 2020 project. Started in 2020 it will allow building a harmonized, probability-based web panel in twelve European countries.[40] Of course, many organizations have made this transition already and we can build on the many experiments and experiences from others.

In addition, we will need to keep exploring the possibilities of combining probability and nonprobability samples (Brick, 2014; Chen et al., 2019). This development is already well underway and the direction of what has been termed survey-assisted modeling will require considerable efforts from survey designers and methodologists, as well as from data users (Heeringa, 2017). Indeed, supplemental data from nonprobability sampling can provide better data for modelling, thereby increasing overall data quality.

Current survey methodology is based on a mixture of theories and sciences such as sampling, psychology, communication, linguistics, data science, and information technology (Platek & Särndal, 2001; Japec & Lyberg (forthcoming)). Thus, there is no comprehensive survey theory and methodology, and survey norms have been allowed to develop in different directions. For example, sampling theory was developed in the 1920's and 1930's and manifested by Neyman's 1934 landmark paper. Fisher (1925) developed randomization principles and together these accomplishments formed modern survey thinking.

Theory developed almost 100 years ago may well need some revision and update. Increasingly high levels of nonresponse in some contexts threaten the viability of probability sampling as a sampling approach. Other error sources, as we have mentioned above (such as measurement error due to the interviewer), result in the reporting of error margins that might be too narrow (i.e., based solely on sampling error), hence these margins tend to be understated. This problem can only be addressed with a vigorous promotion of current best methods, quality control and quality assessments, as well as an increased transparency regarding survey quality.

The increasing issues with probability sampling have led to a renewed interest in nonprobability sampling (Baker et al., 2013; MacInnis et al., 2018) and opt-in panels (Baker et al., 2010; Wang et al., 2015) as well as Bayesian inference (Gelman et al., 2013) as a contrast to the frequentist theory and design-based inference with which most survey researchers are familiar. There are also developments in data collection methods that focus on an increase in data sources. The advent of smartphones and other e-devices, mixed-mode approaches, web panels and administrative records offers opportunities but comes with new, often complicated error sources (Revilla, Ochoa, & Toninelli, 2016).

We also have access to large volumes of inexpensive Big Data that have not been collected for survey purposes but might still be very useful in a survey context, not the least in a 3MC context

[40] See https://www.europeansocialsurvey.org/about/singlenew.html?a=/about/news/essnews0079.html.

(Japec et al., 2015). While the definition of Big Data is currently in flux, there is general agreement that these are complex datasets which cannot be handled by traditional analytical tools, include a variety of sources of content, are available in near-real time (Callegaro & Yang, 2017). For example, data from sources such as Twitter and Facebook, blogs, pictures, videos and internet searches, often more specifically referred to as Big Social Data (Schober et al., 2016), have been explored as a way to gain information to supplement survey data. There are studies where these types of data have been found to be highly correlated with results from traditional surveys such as consumer sentiment studies (see O'Connor et al., 2010 and Daas & Puts, 2014). There are also examples of cases where these types of data have failed to consistently provide an accurate picture of a current event (e.g., Google Flu Trends predictions of the flu in the U.S.) (Butler, 2013). Schober et al. (2016) point out that there are key differences in how participants, survey respondents and individuals making social media postings, actually understand the activities that they engage in. Such perceptional differences among persons hence affect the data and the types of inference that can be made based on these data.

From a 3MC perspective, Big Data can differ extensively among countries and regions, also reflecting the survey conditions faced by 3MC survey staff. Matters in Africa, Latin America and parts of Asia differ substantially from those in Western Europe and North America. Hence, data access and the possibility of analysis are often restricted to specific contexts or local issues. Several Big Data studies are indicative of societal issues, see for example UN Global Pulse (www.globalpulse.org). Nevertheless, there is still an apparent lack of continuity of these studies – most of them appear to be local and event-driven, caused by specific contemporary and often transient societal topics, not completely approaching the target parameter of interest in traditional survey research.

The issue of actually using big data is not straightforward when it comes to survey design or inference. Among the several concerns pointed out by AAPOR in its 2015 Task Force report on Big Data (Japec et al., 2015), the maturity of the data source is crucial. To investigate issues more closely, the first international conference on big data in the social sciences was held in 2018.[41]

One potential use of Big Data in 3MC surveys is to supplement some of the survey questions with social media data. For example, ESS is carried out every two years and it is a very costly survey. Social media data that are highly correlated with some of the attitude questions asked in the ESS such as questions on politics, education and immigration might be collected between rounds. In the long run it seems very unrealistic that 3MC surveys can continue without taking new data sources of an organic[42] nature and associated methodologies into account.

[41] See https://www.bigsurv18.org/program2018 for more information about the conference; see also information regarding the *Social Science Computer Review* special issue on Big Data and Survey Science (https://journals.sagepub.com/doi/full/10.1177/0894439319883393?casa_token=CHtsB6RSEfUAAAAA:5LWImtoxuFXZFtPi1vqo3But6pRv8Dwa0ERFAeIqDhIQRXKF1OAltOL37mpuyCp578fF2eQh_Q4).

[42] Organic data is a term coined by Groves (2011): "Collectively, society is assembling data on massive amounts of its behaviors. Indeed, if you think of these processes as an ecosystem, the ecosystem is self-measuring in increasingly broad scope. We might label these data as ''organic,'' a now-natural feature of this ecosystem." p.868.

6. Summary and recommendations

Section 6 provides a report summary and integrated high-level recommendations including a discussion and justification for the development of a new 3MC survey research discipline.

Comparability is the very purpose and at the same time the main challenge of 3MC survey research. Comparability across languages, cultures, regions and countries, and over time is the issue that distinguishes 3MC surveys from national, monolingual surveys, although one could argue that every survey entails the potential for comparison between groups (the young and the old, the rich and the poor, the rural and the urban) who may have different response styles and may differ in response behaviors. These challenges grow for 3MC surveys comprising several populations and languages, where between-group differences in populations (e.g., different countries) are likely to be larger than between-group differences within populations.

All surveys should try to minimize total survey error, but 3MC surveys also have an obligation to minimize variations across populations' total survey error components. This brings about a number of crucial challenges to surveys sponsors, survey providers, and survey users. Survey sponsors should be informed about how 3MC surveys should be designed and implemented so that comparability is maximized. Design should be informed by study research objectives, best practices, available funds, national expertise and competence, local context, and national survey practices. Survey providers should be able to implement high-quality surveys and be aware of the importance of standardized procedures along with an appropriate degree of localization, to account for cultural variations. Users should be aware of what can and what cannot be accomplished with 3MC surveys and be able to assess the quality and comparability of the data they use and their fitness for purpose.

The quality of survey data depends to a certain extent on users' research purposes. For instance, if the main purpose of a study is to compare trends over time across different countries, or to compare groups within countries, differences between countries in factors such as response rates or acquiescence may be less important than when point estimates are compared. Roger Jowell (1998) strongly advised against using cross-national survey results for ranking purposes. He also warned against interpreting survey data for a country about which little or nothing is known and comparing too many countries simultaneously. These recommendations are especially valuable today, given ease of access to 3MC data and powerful statistical software.

Comparability is a function of the number of populations involved (countries, regions, cultural groups, and so on) and time points covered. When the number of compared populations increases, so, too, do problems with varying competence, resources, control efforts, perceptions of concepts, and so on. Similarly, greater complexity comes with longitudinal surveys spanning decades, or covering instances of radical social change. In light of this insight, 3MC studies should aim to limit the number of populations with a manageable implementation and quality control operation, taking also into consideration financial and research infrastructure resources. Yet there is a countervailing pressure: to include as many populations, e.g., countries, as possible to make the survey more interesting to sponsors and users.

There is no single solution to this tension, to fit all 3MC survey research. However, there are guiding principles, such as Jowell's (1998) rules (see Lyberg et al., 2019, p. 1067), and the recommendations this task force has developed, that can be reshaped into criteria for deciding on the scope of data collection and the conditions that survey design, implementation and documentation should meet to reach levels of comparability that facilitate transparent, ethical and valid knowledge production. It is important that neither data producers nor users ignore error sources, that current best practices are known and applied, and that stakeholders are informed about the quality of the survey data and any resulting limitations.

The recommendations we make below consider both 3MC surveys as such, and developments of 3MC research infrastructure. The latter is essential for promoting interdisciplinary dialogue and cooperation to achieve and implement common standards for quality of 3MC survey data:

- Efforts need to be made to begin reconciliation and unification of the terminology employed to discuss equivalence and comparability across the many disciplines engaged in and contributing to 3MC research.
- To achieve this, efforts to foster interdisciplinary research and collaboration, including training courses are needed. Coordination across projects and organizations in the development of new tools and approaches could greatly accelerate theoretical and methodological developments in 3MC surveys, leading to better quality data and increased efficiencies. This requires dedicated funding. The SERISS initiative in Europe provides an example of how such funding has accelerated and advanced the science and practice of 3MC survey research.
- Breaking down disciplinary barriers also calls for cooperation at both individual and organizational levels. Organizations like AAPOR WAPOR, ESRA, and initiatives such as CSDI, and the methodology-oriented research committees of the American and International Sociological Associations, American and International Political Science Associations, and other stakeholders should form a committee or committees to:

    (i) develop strategies to compile and disseminate information about existing resources and best practices in 3MC survey research, including those listed in the Executive Summary.
    (ii) advance the tools, resources and research in priority areas for future research, such as those outlined in the Executive Summary.
    (iii) develop an interdisciplinary training curriculum that would prepare a new generation of specialists in 3MC survey research.

- 3MC surveys should be designed and implemented taking into account current best practices discussed in this report and summarized in the Executive Summary. The term "current" is key, however; when new knowledge is gained, best practices should be revised and promoted.
- Ongoing 3MC surveys should be reviewed on a regular basis. Both internal and external quality reviews are recommended because they offer different perspectives and new ideas for continuous quality improvement.
- 3MC survey research should be established as a discipline of its own.

This last recommendation demands special justification, since it is critical for the advancement of the science of 3MC research.

Given that 3MC surveys are currently conducted by organizations with varying research traditions and experiences regarding survey quality in general, and 3MC survey quality in particular, this report might have a limited effect in some disciplines that are not familiar with AAPOR/WAPOR activities. Frankly, the field of 3MC research is very large with limited collaboration across different research traditions. For example, while theoretical advances in comparative research are made in specific disciplines, including cultural psychology, cultural sociology, linguistics, organizational science, survey methodology, and psychometrics, both the integration and cross-fertilization of these advances with the aim of improving survey data comparability have been limited. While 3MC surveys share the common goal of producing comparable data across many cultures and countries, the lack of communication and coordination among 3MC survey networks as well as between these networks and researchers has hindered opportunities for advancement in improvements to data quality.

Important progress has been made through the CSDI. CSDI was founded in 2002 with the goal of improving the comparability of survey data across diverse populations. Annual workshops since 2002 have provided a unique forum for researchers from around the world to present and discuss their research related to comparative survey methods. Other initiatives generated by the CSDI executive committee include two large international conferences on Survey Methods in 3MC Contexts with a resulting monograph in 2010 (Harkness et al., 2010b) that won the 2013 AAPOR book award and another monograph in 2019 (Johnson et al., 2019b). A comprehensive online free resource on Cross-Cultural Survey Guidelines and a series of short online courses on international survey research were also produced by members of CSDI (Survey Research Center, 2016; Center for Capacity Building in Survey Methods and Statistics, 2018).

The momentum created by CSDI also led 3MC research to be recognized as an important topic by major national and international organizations. Both the National Center for Education Statistics (NCES) and the Organization for Economic and Co-operation and Development (OECD) have organized seminars in the past two years revolving around the challenges of 3MC surveys (Behr & Zabal (2019) on translation).[43] Moreover, AAPOR now has a session stream labeled 3MC in its annual meeting and a 3MC affinity group. The European initiative called SERISS, mentioned above, was formed to bring together European 3MC survey networks, with funding from the EU's Horizon 2020 research program.

These past and current initiatives have helped foster a growing research network centered around the challenge of comparability in 3MC surveys and have helped develop a collective research literature. However, much more remains to be done to engage additional 3MC survey networks, increase connections with researchers conducting cross-cultural research in other fields, particularly in new disciplinary fields such as computational linguistics. A funded effort to increase communication and foster interdisciplinary research and collaboration is urgently needed to advance the science and practice of 3MC survey research.

[43] See also http://www.oecd.org/skills/piaac/events/.

Further, in order to develop the field, we need to make 3MC research a discipline of its own. So, what does such a development entail? According to Groves (2018), a number of criteria must be fulfilled before a field can declare itself a discipline. The following list is one possible set of such criteria.

f. an academic curriculum should be developed;
g. a professional organization should be created;
h. a scientific journal or a named set of publication outlets should be available to the discipline;
i. the discipline should have a common set of shared values and research principles; and
j. there should be deep ongoing work in knowledge domains.

We cannot yet claim that all these criteria have been fulfilled. There are a few informal interest groups with CSDI at the forefront, research papers are presented at many conferences, and research papers are published in journals that normally cover topics from official statistics to ethnology. Deep ongoing work is indeed being done, but there are problems with outreach across this large field and the diffusion of innovations across disciplines and countries is uneven at best.

According to Groves (2019), all fields need people, people that can be replaced over time. For a field to become a discipline it has to be large enough to attract a critical number of students, faculty, and practitioners. The 3MC literature is comprised of a number of monographs and resources that already now serve as teaching material. What is lacking is a systematic training program, including textbooks for undergraduate and graduate levels. Today scattered single courses are taught in universities, but to move to a product that provides an academic certification, an integration of courses is needed. Also, there need to be jobs within the discipline area and here, there appears to be no shortage of opportunities. However, there must be a structured process for training new generations for the field to develop further.

Members of the 3MC field should formalize existing informal groups, form a professional group, and develop this discipline focusing on the criteria above. A group of members selected from this Task Force are in the initiation stages of this process.

Appendix 1 – Task Force Charge

May 17, 2018

Mission for an AAPOR/WAPOR task force on quality of comparative surveys

Background

Comparative surveys are surveys that study more than one population with the purpose to compare various characteristics of the populations. To achieve comparability these surveys need to be carefully designed according to state-of-the-art principles and standards.

Examples of comparative surveys are the European Social Survey, the International Social Survey Program, the Gallup World Poll, the World Values Survey, the European Union Survey of Income and Living Conditions, the Programme for International Student Assessment, and the Global Marketing Compensation Survey.

Some researchers use the notion Surveys in Multinational, Multiregional, and Multicultural Contexts (3MC surveys) as an alternative to comparative surveys. In some research groups the acronym 3MC has become almost a trade-mark.

There are a myriad of 3MC surveys that are conducted within the official statistics, academic, and private sectors. They have in general become increasingly important to global and regional decision-making as well as theory-building. At the same time these surveys display considerable variation regarding methodological and administrative resources available, organizational infrastructure, awareness of error sources and error structures, level of standardized implementation across populations, as well as user involvement. These circumstances make 3MC surveys vulnerable from a quality perspective. Quality problems present in single-population surveys are magnified in 3MC surveys and new quality problems specific to 3MC surveys must be added on top of the former.

We believe that so far quality problems have not been well handled in most 3MC surveys. The substantive output in terms of actual comparisons of populations (often countries) is often rather impressive. The output could be league or ranking tables, research reports, and assessment analyses. This wealth of output is, however, usually not accompanied by a corresponding interest in informing researchers, decision-makers, and other users about quality shortcomings. This can lead to understated margins of error and league tables that appear more precise than they actually are. There are also cases where researchers are informed about quality shortcomings but opt to ignore those in their research reports. There are of course many possible explanations for this state of affairs. One is that 3MC surveys are very expensive and the formidable planning and implementation leaves relatively little room for a comprehensive treatment of quality issues. Another explanation is that the survey-taking cultures among survey professionals vary considerably across nations as manifested by varying degrees of methodological capacity, risk assessment, and willingness to adhere to specifications that are not normally applied.

The literature on data quality in 3MC surveys is scarce compared to the substantive literature. There are exceptions, though, including the Cross-Cultural Survey Guidelines developed by the University of Michigan and members of the International Workshop on Comparative Survey Design and Implementation (CSDI). AAPOR has created a cross-cultural and multilingual research affinity group and some 3MC surveys have advanced continuing data quality research programs. Recently OECD organized a seminar on interviewer errors in their Programme for the International Assessment of Adult Competencies. In May 2018 OECD will organize another seminar, this time on translation issues. Members of the CSDI Workshop have produced three monographs that treat advances in the field of 3MC surveys. There are also scattered book chapters and journal articles that discuss 3MC and quality.

Despite these efforts we believe that there is need for a task force that can address the most pressing challenges concerning data quality in this field. The output will be a set of recommendations regarding quality issues backed by justifications.

The scope

The task force will investigate, discuss, and comment upon aspects of the following areas:

A. What's so special about 3MC surveys?

This first point will highlight some important features of 3MC surveys and how they differ from those in single-population surveys. Issues that will be discussed include the various meanings of equivalence, the need for an infrastructure that can handle the methodological and administrative challenges involved, the fact that populations such as nations and cultural subpopulations can be very different on several dimensions, that special error sources such as translation of survey materials and adaptation of questions in the source questionnaires are present, and that the research traditions vary a lot. For instance, in assessment surveys much energy goes to psychometric considerations rather than to survey errors.

B. The notion of quality in a 3MC setting

This second point is a general assessment of how quality should be perceived in a 3MC setting. There are a number of quality frameworks in survey science that can help identify error sources and provide guidance regarding their effects on estimates. There are basically two types of frameworks, namely total survey error (TSE) frameworks and frameworks that combine different quality dimensions, both quantitative and qualitative. TSE frameworks in 3MC surveys are discussed in Smith (2011), Pennell et al (2016), and Lyberg et al (2018). Smith, for instance, introduces the concept "comparison error", which implies that TSE in 3MC settings is not only a matter of quality of estimates but also how well these estimates live up to requirements regarding equivalence, i.e., how comparable the estimates are across populations. Quality in 3MC surveys is also a matter of qualitative dimensions such as timeliness, relevance, accessibility, and whether comparisons are at all possible for some subsets of populations due to large differences in perceptions of concepts and general capacity.

We intend to discuss how quality should be perceived and conveyed to the users in a 3MC setting and whether it is possible to define criteria and indicators for good quality. Good quality is ideally decided together with the users but in 3MC surveys the distance between producer and user is very large and the producer should assume a greater responsibility for quality than is normally the case in single-population surveys.

C. Basic design and implementation

This third point provides an overview of the state-of-the-art regarding design and implementation issues given the important features mentioned above. Conducting a 3MC survey is a huge undertaking. If the number of populations is very large it is extremely hard to make valid comparisons and also to manage the entire operation. We intend to discuss whether there are upper limits to consider here. How should resources be allocated given different scenarios? Are there scenarios that cannot possibly generate outputs of good quality and, if so, can they be readjusted?

All 3MC surveys have a process for design and implementation. Some of them will be described given their levels of ambition. Sometimes ambition levels are out of step with the quality provided. In some surveys almost everything about design and implementation is allowed to vary across populations. It is not uncommon that, say, participating countries are asked to administer a source questionnaire and deliver the results to a central site. Others allow a great deal of freedom on the part of the local data collection organizations and rely on what is called output harmonization to adjust for this variation. But if the freedom is such that the basic requirement of approximately similar essential survey conditions is not fulfilled comparisons become difficult. Some surveys use rather extensive input harmonization, i.e., a set of specifications on how design steps should be carried out, but without quality control local organizations might deviate from the specifications, which can have an effect on quality. Very ambitious surveys use input harmonization and check if specifications are adhered to. The problem is that often these checks come too late to be able to impact the data collection.

Lack of a timely quality control is a problem in most, if not all, 3MC surveys. Even if specifications exist, they might not be understood, affordable, in line with local best practices, or they might be downright overwhelming. Here we intend to discuss this problem and some others that are related to quality control, such as data fabrication, lack of know-how, and limited appreciation for some types of error both among producers and advanced users.

Report and Recommendations

Most recent task force reports have treated new or emerging topics such as the use of big data, nonprobability sampling, and mobile technology. This task force report will discuss issues regarding multipopulation surveys that are already an established part of survey science. The reason is that we believe that many of these surveys are, in a broad sense, problematic from a quality perspective and need improvements. The report will end with a number of recommendations.

The recommendations will concern 3MC surveys that in most cases are already up and running and cover among other things the following:

- A unified view on the definition of quality
- The importance of a solid infrastructure and a central management team
- Error structures
- Examples of good quality assurance and quality control procedures used in practice
- Quality reporting and the role of the user
- New technology and a changing survey landscape
- Urgent areas for improvement
- Prevailing issues and ideas for the future

AAPOR/WAPOR Task Force (TF) on Comparative Surveys

Preliminary Time Schedule

No	Time periods	Task action list
1	March 2018	Finalize draft of task force charge, presented at CSDI
2	April-May 2018	Team members assigned to the TF including satellite members (STF)
3	April 4 2018	Task force charge sent to AAPOR council for approval and comments
4	April 24 2018	Task force charge revised and sent to WAPOR for comments
5	May 16, 2018	Meeting at AAPOR. Discussion of TF charge, time schedule, work format, and possible sponsors of workshop. Assigning responsibilities and writing tasks to each team member.
6	June 15 2018	Draft extended outline due
7	July 20 2018	First rough draft (4 pages per section) due
8	July 30-August 1, 2018	TF members that are available meet and produce first draft
9	September 15, 2018	Comments on first draft from all TF and STF members due
10	September 16, 2018 -October 31, 2018	TF works on second draft
11	November 1 -30 2018	Assembling comments on second draft from all stakeholders including AAPOR and WAPOR councils. Possible workshop with TF and available STF members pending sponsor support

12	**December 1, 2018 - January 15 2019**	**TF works on third draft**
13	January 16, 2019 - February 15, 2019	Assembling comments on third draft from all stakeholders
14	February 16, 2019 - March 15, 2019	TF produces final draft, presented at CSDI
15	March 16, 2019 – March 31, 2019	Comments on final draft from all stakeholders
16	April 1, 2019 – April 15, 2019	Final report produced
17	May 2019	Final report presented at AAPOR

Proposed Task Force Membership

1. **AAPOR Members**

- Lars Lyberg (Inizio, Sweden) Co-Chair
- Kristen Cibelli Hibben (University of Michigan)
- Julie de Jong (University of Michigan)
- Timothy Johnson (University of Illinois at Chicago)
- Michael Robbins (Princeton University)
- Tom Smith (NORC at the University of Chicago)
- Ineke Stoop (The Netherlands Institute for Social Research)
- Mandy Sha (Cross-Cultural and Multilingual Research Affinity Group)

2. **WAPOR Members**

- Beth-Ellen Pennell (University of Michigan) Co-Chair
- Irina Tomescu- Dubrow (Institute of Philosophy and Sociology, Polish Academy of Sciences (PAN) and CONSIRT at The Ohio State University and PAN)
- Linda Guerrero (Social Weather Stations, Philippines)
- Dorothee Behr (Gesis, Germany)
- Jibum Kim (East Asian Social Survey, Korea)
- Elizabeth Zechmeister (LAPOP)

Appendix 2 – Table 2 References

Adams-Esquivel, H. (1991). Conceptual adaptation vs. back-translation of multilingual instruments: How to increase the accuracy and actionability of multilingual surveys. Paper presented at the annual meeting of the American Association for Public Opinion Research; Phoenix, AZ.

Anderson, R.T., Aaronson, N.K. & Wilkin, D. (1993). Critical review of the international assessments of health-related quality of life. *Quality of Life Research* 2: 369-395.

Behling, O. & Law, K.S. (2000). *Translating Questionnaires and Other Research Instruments*. Thousand Oaks, CA: Sage.

Borg, I. & Shye, S. (1995). *Facet Theory: Form and Content*. Thousand Oaks, CA: Sage.

Blumer, M. & Warwick, D.P. (1993). Social Research in Developing Countries: Surveys and Censuses in the Third World. London: Wiley.

Buil, I., de Chernatony, L., & Martinez, E. (2012). Methodological issues in cross-cultural research: An overview and recommendations. *Journal of Targeting, Measurement and Analysis for Marketing* 20: 223-234.

Craig, C.S. & Douglas, S. P. (2000). *International Marketing Research*, Second Edition. Chichester, UK: Wiley.

Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R. & Hausherr, M. (2015). The comparability of measurements of attitudes toward immigration in the European Social Survey: Exact versus approximate measurement equivalence. *Public Opinion Quarterly* 79: 244-266.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology* 40: 55-75.

de Jong, J., Dorer, B., Lee, S., Yan, T., & Villar, A. (2019). Overview of questionnaire design and testing. In T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, & B. Dorer (Eds.) *Advances in Comparative Survey Methods* (pp. 115-137). Hoboken, NJ: Wiley.

Devins, G.M., Beiser, M, Dion, R., Pelletier, L.G., and Edwards, R.G. (1997). Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translation of the affect balance scale. *American Journal of Public Health* 8: 794-799.

Dressler, W.M., Viteri, F.E., Chavez, A., Frell, G.A.C. & Dos Santos, J.E. (1991). Comparative research in social epidemiology: Measurement issues. *Ethnicity and Disease* 1: 379-393.

Dunnigan, T., McNall, M. & Mortimer, J.T. (1993). The problem of metaphorical nonequivalence in cross-cultural survey research. *Journal of Cross-Cultural Psychology* 20: 133-151.

Eckensberger, L.H. (1973). Methodological issues of cross-cultural research in developmental psychology. In J.R. Nesselroade & H.W. Reese (Eds.) *Life-Span Developmental Psychology: Methodological Issues* (pp. 43-64). New York: Academic Press.

Elder, J.W. (1973). Problems of cross-cultural methodology: Instrumentation and interviewing in India. In M. Armer & A.D. Grimshaw (Eds.) *Comparative Social Research: Methodological Problems and Strategies* (pp. 119-144). New York: Wiley.

Elder, J.W. (1976). Comparative cross-national methodology. In A. Inkeles, J. Coleman & N. Smelser (Eds.), *Annual Review of Sociology*, Volume 2 (pp. 209-230). Palo Alto, CA: Annual Reviews, Inc.

Ellis, B.B., Minsel B. & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. *International Journal of Psychology* 18: 665-684.

Eyton, J. & Neuwirth, G. (1984). Cross-cultural validity: Ethnocentrism in health studies with special reference to the Vietnamese. *Social Science and Medicine* 5: 447-453,

Feldkircher, M. (1998). Religious orientations and church attendance. In J.W. van Deth (Ed.), *Comparative Politics: The Problem of Equivalence* (pp. 86-110). London: Routledge.

Flaherty, J.A., Garviria, M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A. & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous and Mental Disease* 176: 257-263.

Frey, F.W. (1970). Cross-cultural survey research in political science. In R.T. Holt & J.E. Turner (Eds.) *The Methodology of Comparative Research* (pp. 173-294). New York: Free Press.

Herdman, M., Fox-Rushby, J. & Badia, X. (1997). 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research* 6: 237-247.

Hox, J.J., de Leeuw, E., & Brinkhuis, M. (2010). Analysis models for comparative surveys. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.Ph. Mohler, B.-E. Pennell & T.W. Smith (Eds.) *Survey Methods in Multinational, Multiregional and Multicultural Contexts* (pp. 395-418). Hoboken, NJ: Wiley.

Hox, J.J., de Leeuw, E.D. & Zijlmans, E.A.O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology* 6: 87.

Hui, C.H. & Traindis, H.C. (1983). Multistrategy approach to cross-cultural research: The case of locus of control. *Journal of Cross-Cultural Psychology* 14: 65-83.

Iyengar, S. (1976). Assessing linguistic equivalence in multilingual surveys. *Comparative Politics* 8(4): 577-589.

Johnson, T.P. (1998). Approaches to establishing equivalence in cross-cultural and cross-national survey research. *Cross-Cultural Survey Equivalence*. Mannheim, Germany: ZUMA-Nachrichten Spezial, 3: 1-40.

Johnson, T.P., Pennell, B.-E., Stoop, I.A.L. & Dorer, B. (2019). The promise and challenge of 3MC research. In T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, & B. Dorer (Eds.) *Advances in Comparative Survey Methods* (pp. 3-12). Hoboken, NJ: Wiley.

Kashgary, A.D. (2011). The paradox of translating the untranslatable: Equivalence vs. non-equivalence in translating from Arabic into English. *Journal of King Saud University – Languages and Translation* 23:47-57.

Kenny, D. (2001). Equivalence. In M. Baker (Ed.) Routledge Encyclopedia of Translation Studies (pp. 77-80). London: Routledge.

Kleiner, B., Pan, Y., & Bouic, J. (2009). The impact of instructions on survey translation: An experimental study. *Survey Research Methods* 3(3): 113-122.

Kohn, M.I. & Słomczyński K.M. (1990). *Social Structure and Self-Direction: A Comparative Analysis of the United States and Poland*. Cambridge, MA: Basil Blackwell.

Kuechler, M. (1987). The utility of surveys for cross-national research. *Social Science Research* 16: 229-244.

Leonardi, V. (2000). Equivalence in translation: Between myth and reality. *Translation Journal* 4(4).

Padilla, J.-L., Benitez, I., & van de Vijver, F.J.R. (2019). Addressing equivalence and bias in cross-cultural survey research within a mixed methods framework. In T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, & B. Dorer (Eds.) *Advances in Comparative Survey Methods* (pp. 45-64). Hoboken, NJ: Wiley.

Prince, R. & Mombour, W. (1967). A technique for improving linguistic equivalence in cross-cultural surveys. *Journal of Social Psychology* 13: 229-237.

Riordan, C.M. & Vandenberg, R.J. (1994). A central question in cross-cultural research. *Journal of Management* 20: 643-671.

Saule, B. & Aisulu, N. (2014). Problems of translation theory and practice: Original and translated text equivalence. *Procedia – Social and Behavioral Sciences* 136: 119-123.

Sechrest, L., Fay, T.I., & Hafeez Zaidi, S.M. (1972). Problems of translation in cross-cultural research. *Journal of Cross-Cultural Psychology* 3: 41-56.

Sekaran, U. (1983). Methodological and theoretical issues and advancements in cross-national research. *Journal of International Business Studies* 14(2): 61-73

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies* 26: 597-619.

Špirk, J. (2009). Anton Popovič's contribution to translation studies. *Target: International Journal of Translation Studies* 21(1): 3-29.

Stevelink, S.A.M. & van Brakel, W.H. (2013). The cross-cultural equivalence of participation instruments: A systematic review. *Disability & Rehabilitation* 35(15): 1256-1268.

Teune, H. (1977). Analysis and interpretation in cross-national survey research. In A. Szalai & R. Petrella (Eds.) *Cross-National Comparative Survey Research: Theory and Practice* (pp. 95-128). Oxford: Pergamon.

Teune, H. (1990). Comparing countries: Lessons learned. In *Comparative Methodology: Theory and Practice in International Social Research* (pp. 38-62). London: Sage.

Triandis, H.C. (1972). The Analysis of Subjective Culture. New York: Wiley.

Tsai, T.-I., Luck, L., Jefferies, D., et al. (2018). Challenges in adapting a survey: Ensuring cross-cultural equivalence. *Nurse Researcher* 26(1): 28-32.

van de Vijver, F. & Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks, CA: Sage.

van Deth, J.W. (1998). *Comparative Politics: The Problem of Equivalence*. London: Routledge.

van Herk, H. (2000). Equivalence in a Cross-National Context: Methodological & Empirical Issues in Marketing Research. PhD Dissertation, Katholieke Universiteit Brabant, The Netherlands. Accessed at: https://pure.uvt.nl/ws/portalfiles/portal/358466/82801.pdf

van Herk, H., Poortinga, Y.H. & Verhallen, T.M.M. (2005). Equivalence of survey data: Relevance for international marketing. *European Journal of Marketing* 39: 351-364.

Verba, S., Nie, N.H. & Kim, J.-O. (1978). *Participation and Political Equality: A Seven-Nation Comparison*. Cambridge: Cambridge University Press.

Veselinova, D. (2014). Teoretical discutions about equivalence in translations. *European Scientific Journal*. Accessed at: https://pdfs.semanticscholar.org/63ce/1cc8467b162fc2b4902352c8223c0f8725ef.pdf

Appendix 3 – Smith's 2011 TSE and comparison error figure

Figure A

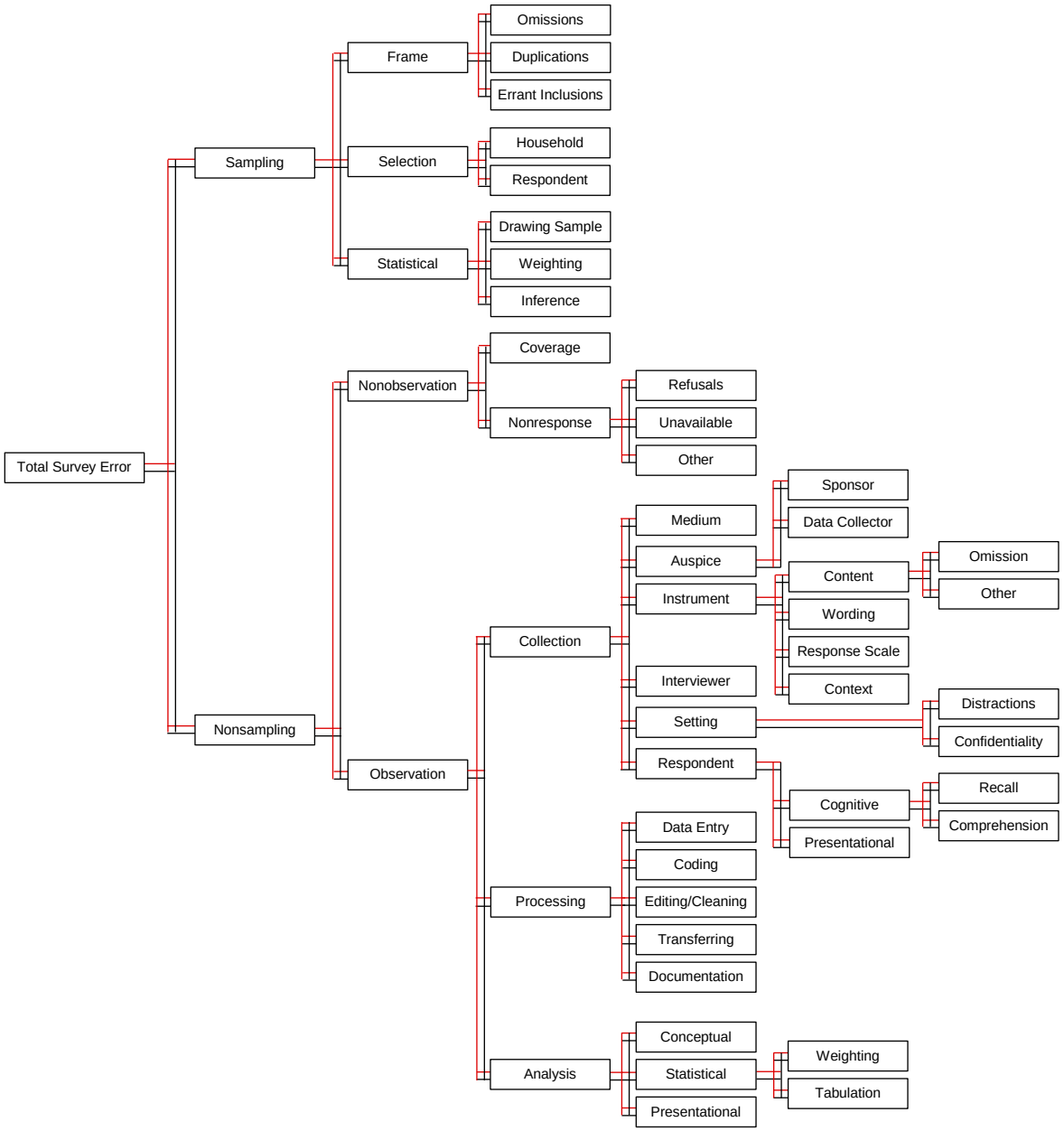
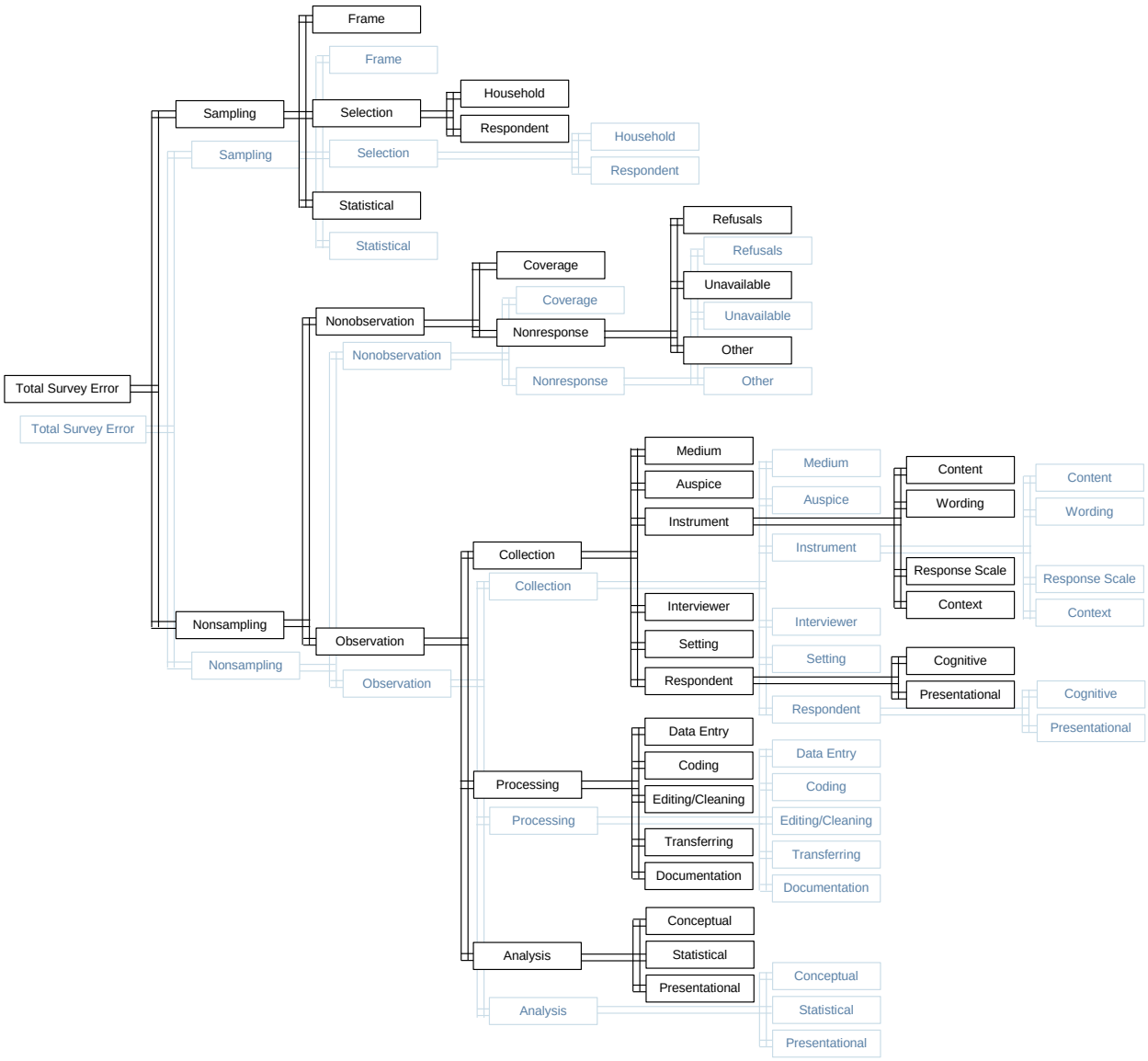


Figure B


Total Survey Error
Total Survey Error
Sampling
Sampling
Nonsampling
Nonsampling
Frame
Frame
Selection
Selection
Statistical
Statistical
Household
Respondent
Household
Respondent
Nonobservation
Nonobservation
Observation
Observation
Coverage
Coverage
Nonresponse
Nonresponse
Refusals
Refusals
Unavailable
Unavailable
Other
Other
Collection
Collection
Processing
Processing
Analysis
Analysis
Medium
Auspice
Instrument
Interviewer
Setting
Respondent
Medium
Auspice
Instrument
Interviewer
Setting
Respondent
Content
Wording
Response Scale
Context
Cognitive
Presentational
Content
Wording
Response Scale
Context
Cognitive
Presentational
Data Entry
Coding
Editing/Cleaning
Transferring
Documentation
Data Entry
Coding
Editing/Cleaning
Transferring
Documentation
Conceptual
Statistical
Presentational
Conceptual
Statistical
Presentational

Appendix 4 – Pennell et al. 2017 TSE framework adapted for 3MC surveys

Figure A: TSE Representation in a cross-cultural context

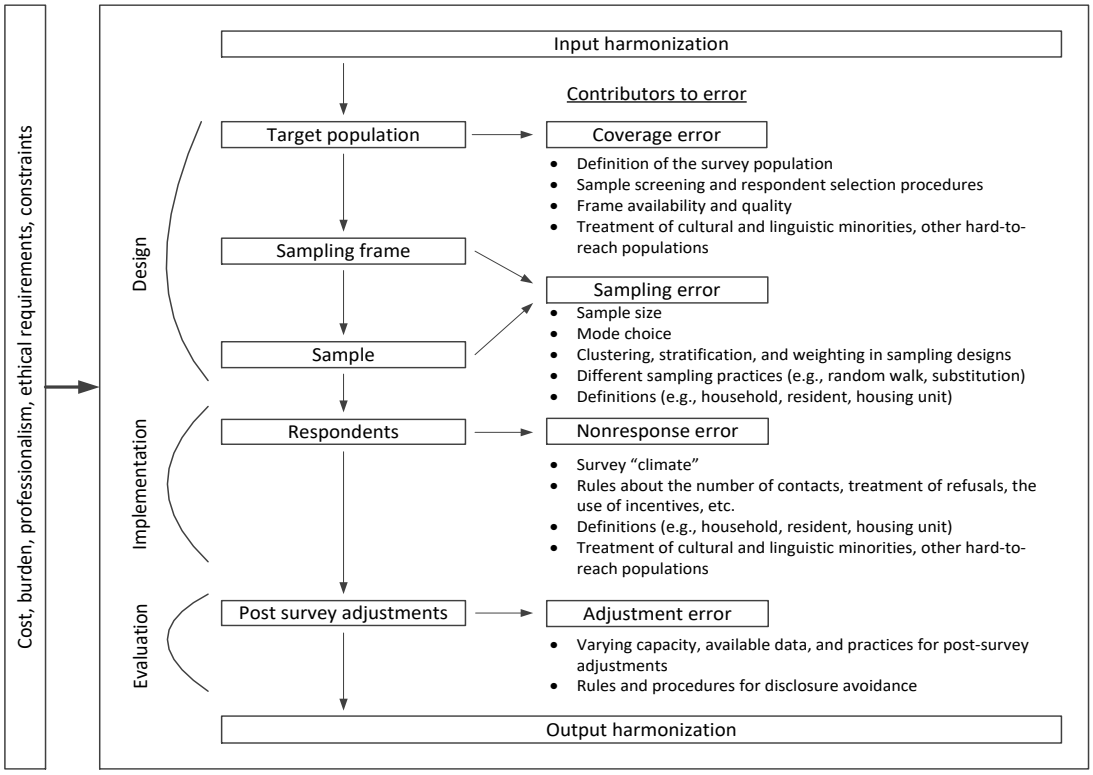
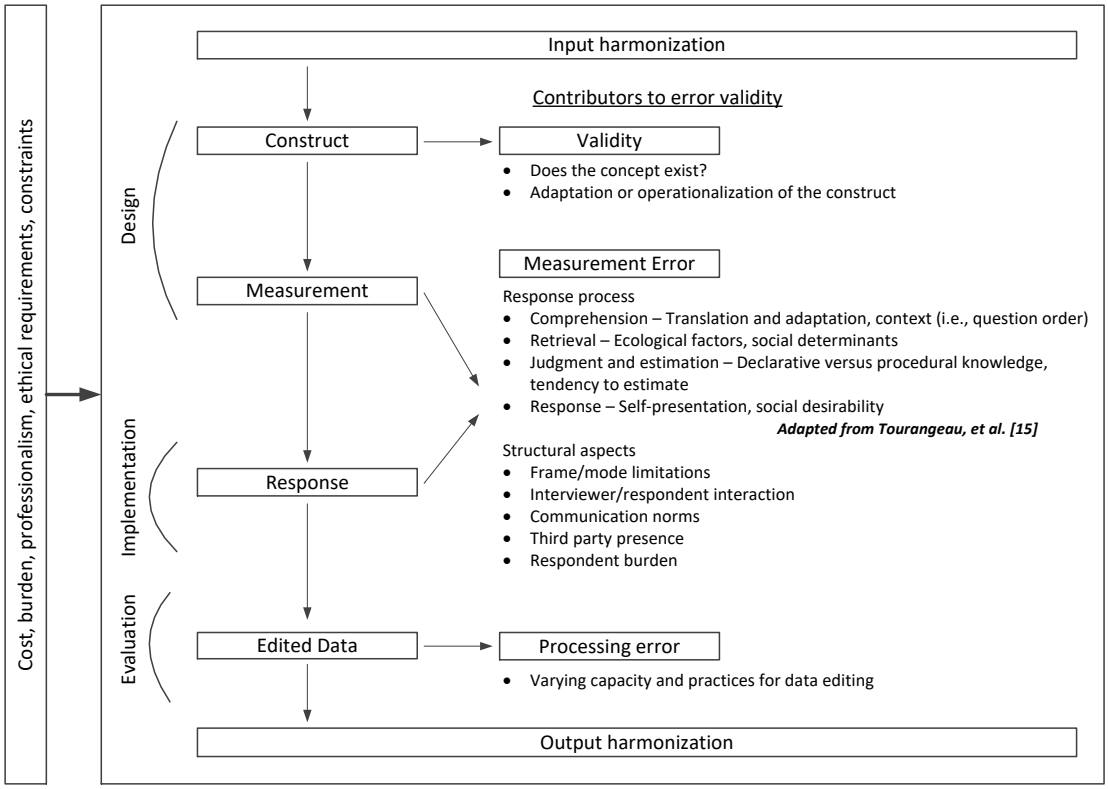


Figure B: TSE Measurement in a cross-cultural context



Adapted from Groves, et al. [14]

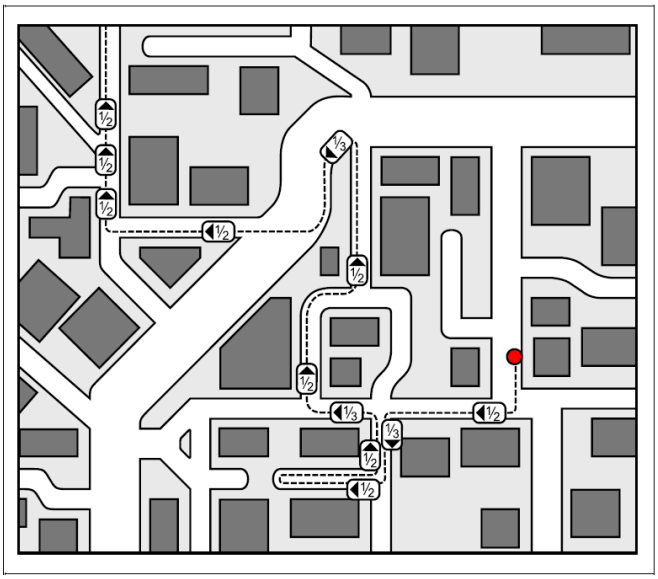
Appendix 5 – Bauer's random route alternatives

True Random Route (TRR)

For each starting address, an interviewer is provided with an n x 3 direction matrix, where n is the number of rows of the matrix and represents the nth junction on the random route. The directions on the grid in each row (left, right, straight) are randomly assigned by the central sampling team from the master set of six combinations: L-R-S, L-S-R, R-L-S, R-S-L, S-L-R, S-R-L.

At the first junction, the interviewer should follow the direction in Row 1 and Column 1. If this direction is not available, the interviewer should revert to Column 2 in Row 1. Should the chosen direction be left, the street farthest left has to be selected; if it is right, the street farthest right must be taken. Interviewers mark the direction and proceed (as illustrated by the grey shading) to the next junction on their random route. At the second junction, the interviewer should follow the direction in Row 2 and Column 1, or Column 2 in Row 2 if direction 1 is not available (and so on). This procedure continues, moving to a new row at each junction, until the required number of households along the route is sampled. The map illustrates the direction of the random route based on the direction matrix.

Map showing random walk



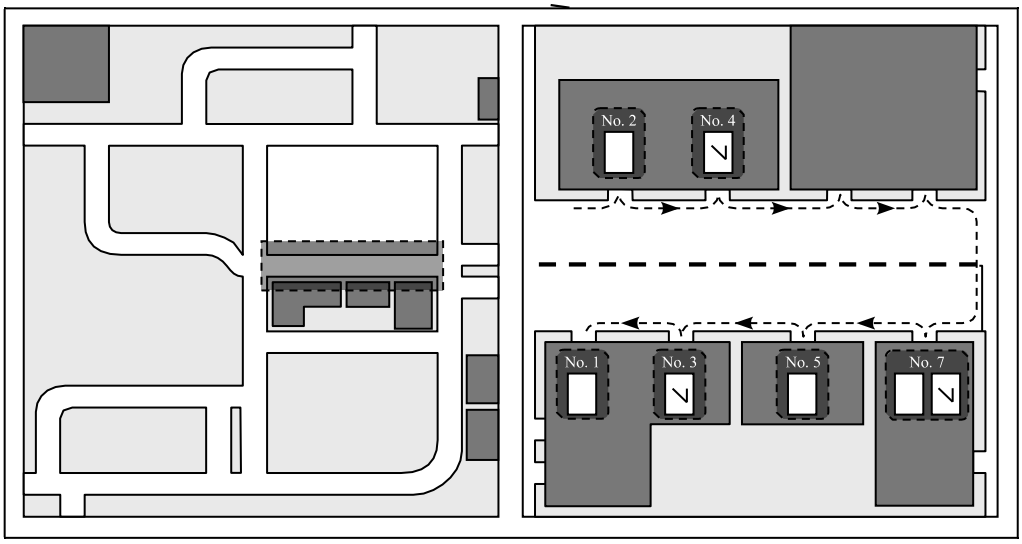
Direction matrix

	1	2	3
1	right	left	straight
2	left	straight	right
3	left	right	straight
4	straight	left	right
5	left	right	straight
6	right	straight	left
7	right	left	straight
8	left	right	straight
9	left	right	straight
10	straight	right	left
11	straight	left	right
12	right	straight	left
	…	…	…

Street Section Sampling (SSS)

The second approach tackles the problem of unequal start locations and removes the drawback of random routes in being a non-probability sample with unknown selection probabilities. This method is based on the application of increasingly accurate and accessible road map data. Geographic data file providers like HERE, TomTom and Auto-motive Navigation Data (AND) enable the usage of international map data. Cities and land registry offices have up-to-date regional road maps and the open source project OpenStreetMap (OSM) provides free access to high quality data (Ciepłuch et al. 2010, Girres & Touya, 2010, Haklay 2010).

Example of a street section sample. Random selection of a street section in which every 3rd household is questioned. The first contact is randomly chosen from the first three households.



For an interviewer, the selection of households would typically look like the example above. The left panel shows the selection of a street section, the right panel how an interviewer selects households by contacting every 3rd one. Beginning from the start crossroad at the specified street side, the interviewer walks in the direction of the other end of the street section. The interviewer counts households until reaching the assigned start household in the street section, the first contact. From there, every 3rd household is contacted and interviewed. When the interviewer reaches the end of the street section, s/he changes to the other side and proceeds until returning to the starting crossroad. After the interviewer has completed a street section, s/he is assigned to another randomly selected one. This process continues until the required number of households is obtained.

The street section sampling approach is a possible solution for unequal distribution of starting locations. However, it is associated with additional resources for coding and scripting, as well as fieldwork training. Further, there can be street sections with no households that should ideally be identified before the interviewers start fieldwork. It might, therefore, be more practical for survey agencies that already use random route sampling to apply the TRR approach as it follows the basic principles of general random route and only requires changing direction instructions to

reduce bias in household selection. Using longer interviewer routes could further reduce the effect of unequal distribution of starting locations in the TRR samples.

Appendix 6 – 3MC Survey documentation standards for study-level and variable-level metadata and auxiliary data

Study-level metadata

- Principal investigator(s). Principal investigator name(s), and affiliation(s) at time of data collection.
- Title. Official title of the data collection.
- Funding sources, including grant number and related acknowledgments.
- Data collector/producer. Persons or organizations responsible for data collection, and the date and location of data production.
- Project description. Describes the project and its intellectual goals. Indicates how the data articulate with related datasets: is the study part of a cross-national project/ a single-country multi-cultural or multi-ethnic project? Was the survey fielded on its own or part of a larger study? If part of a larger study, provides name (title) of the larger study. Cites publications providing essential information about the project. A brief project history detailing major difficulties faced or decisions made is useful.
- Study design. Describes the study design: e.g. single-cross sectional design/repeated cross-sectional design/panel design/ rotating panel design, and so on.
- Target population, and, if applicable, survey population. Describes the population for which the study aims to make inferences (target population), and if different from target population, the actual population from which the survey sample is drawn (survey population). Description of survey population provides exclusion/inclusion criteria.
- Unit(s) of observation: Who or what was studied.
- Sample and sampling procedures. Describes in detail:
  (i) The sampling frame (primary sampling units, number of sampling frames used; whether sampling frames are new or preexisting; if preexisting, whether they were updated for current survey);
  (ii) The sampling procedures: Random/non-random sample design. Number of selection stages. For random sample, describes type: simple/systematic/stratified/cluster/ matched-pairs sampling/random route with saved address listing of households (for random walk with saved address listing, useful to indicate whether walk separate from or combined with interviewing). If non-random – specify type: random walk with interview and no saving of household addresses; quota sample; snowball sample; expert sample, other. Description of sampling procedures also indicates if more complex methods are used within given sample design type (e.g. if substitution of dropouts was permitted; if clustering was used at more than one selection stage; defining the clustering groups for any given stage; characteristics used for stratification, and so on.). If available, a copy of the original sampling plan should be included as an appendix.
- Incentives for survey participants. Indicates if respondents received incentives related to the study. If yes, specifies if incentive was conditional on completing interview (i.e., when it was handed out), describes incentive(s) and indicates where in the dataset

incentive details are at the respondent level in the event of differential incentive allocation.

- Interview mode(s). Describes interview mode or modes used to collect the study data. Specifies criteria for change in interview mode within survey. Indicates if interview mode was changed within respondent; if yes – specify interview modes used and criteria for switching interview mode.
- Data collection instrument/instrument versions. Provides all questions (full question wording) posed to survey participants, the sequence they were posed in, thus recreating the survey context. Should specify if instrument applied to full sample, or to sub-sample. If the latter, provide sub-sample characteristics. Since computer-assisted data collection modes often do not produce hardcopy instruments—or if they can be generated, they may be difficult to read. Increasingly, survey organizations that use CATI and CAPI systems provide online versions of the entire electronic survey scripts, complete with programming, skip logic, question piping, and related technical features. Also specifies original language(s) of data collection instrument.
- Pretesting of data collection instrument(s) in the language(s) they were originally developed. Indicates whether data collection instrument(s) were pretested. If yes, describes pretesting methodology (qualitative/quantitative/larger pilot study). This documentation item is distinct from reporting on the methodology of pretesting translated instruments.
- Translation of data collection instrument. Details the translation process and how the translation (or translations) was reviewed and assessed. Includes information about the languages into which data collection instrument(s) were translated; the organization(s) or firm(s) used for translation of each of the questionnaires from the source to target languages; composition and skills of the translation and review team (e.g. professional translators (freelancers/ translation agencies), member(s) of the survey project; external content experts; mother tongue of team members; their translation experiences in general, and experience relevant for the translation task (e.g. study topic, experience of translating questionnaires, knowledgeable about questionnaire design); written instructions and guidance from project team (if any). Details of the translation and review approach: double translation; split translation; single translation. Individual reconciliation/review; team reconciliation/review; back translation; rating tasks; pretesting of translated instruments (qualitative/quantitative/larger pilot study).
- Interviewer training. Describes the procedures employed to train interviewers.
- Fieldwork implementation and monitoring. Indicates the number of interviewers and the interviewer assignment process. Indicate all languages fielded in each study site, the process for determining the language of the interview, and any use of on unwritten languages and on-the-fly translation. Describes all procedures for monitoring fieldwork during the data collection process.
- Response rate. Indicates the proportion of sampled units who actually participated in the survey; calculated using the American Association of Public Opinion Research (AAPOR) standard definitions, which have also been adopted by WAPOR (American Association for Public Opinion Research, 2016). For longitudinal studies, the retention rate across waves is also noted.

- Weighting. Indicates if weight variables are available. If yes, provides number of weights, and types of weight variables (e.g. design weights with/without taking into account existing non-response rates and their impact on probabilities/calibrated weights/post-stratification weights/combined weights). Describes methodology for constructing each weight variable (e.g. list variable/variables used to construct post-stratification weights and specify if raking/cross-table/mixed was used to construct the weight) and indicates how the weight variables should be used. Describes any weight trimming implemented.
- Dates and geographic location of data collection, and time period covered.
- Data source(s). If a dataset draws on resources other than surveys, indicates the original sources or documents from which data were obtained.

Variable-level metadata

- *The exact question wording or the exact meaning of the datum.* Sources should be cited for questions drawn from previous surveys or published work.
- *A link between the question and the variable*, by including the question number in the variable label.
- *Universe information, i.e., who was actually asked the question*. Indicates exactly who was asked and was not asked the question. If a filter or skip pattern means that data on the variable were not obtained for all respondents, this information is provided together with other documentation for that variable.
- *Exact meaning of codes (i.e., variable values)*. The documentation shows the interpretation of the codes (i.e., values) assigned to each variable. For some variables, such as occupation or industry, summary descriptions in the codebook and value labels can be complemented with more extended information provided in an appendix.
- *Missing data codes.* Codebook and value labels show the interpretation of codes/values that fall outside of the range of values corresponding to 'valid' answers. Different types of missing data (e.g. don't know; refused to answer; not asked) should have distinct codes/values. These codes should be used systematically across the entire dataset(s).
- *Unweighted frequency distribution or summary statistics*. These distributions should show both valid and missing cases.
- *Imputation and editing information*. Documentation identifies data that have been estimated or extensively edited.
- *Cumulative scaling.* Codebook should indicate whether a set of variables is designed to conform to a cumulative ordering process (i.e., Guttman scale).
- *Details on constructed variables.* For variables constructed using other variables, documentation should include "audit trails", indicating exactly how they were constructed, what decisions were made about imputations, and the like. Ideally, documentation would include the exact programming statements used to construct such variables.
- *Details on weight variables.* The construction of each type of weight variable (e.g. design weights with or without taking into account existing non-response rates and their impact

on probabilities; calibrated weights; post-stratification weights; combined weights) in the dataset needs to be described in detail in the codebook. For example, what variable(s) were used to construct post-stratification weights; what is the source (or sources) of those variables; what was the method - raking/cross-table/mixed? How many cells of particular variables were used in the process of constructing post-stratification weight/weights? If combined weights are available – which weight variables were combined and how (e.g. multiplication, other - describe).

- *Location in the data file*. For raw data files, documentation should provide the field or column location and the record number (if there is more than one record per case). If a dataset is in a software-specific system format, location is not important, but the order of the variables is. Ordinarily, the order of variables in the documentation will be the same as in the file; if not, the position of the variable within the file must be indicated.
- *Variable groupings*. For large datasets, it might be useful to categorize variables into conceptual groupings.

Ancillary information

- *Interviewer guide*. Details on how interviews were administered, including probes, interviewer specifications, use of visual aids such as hand cards, and the like.
- *Flowchart of the data collection instrument.* A graphical guide to the data, showing which respondents were asked which questions and how various items link to each other. This is particularly useful for complex questionnaires or when no hardcopy questionnaire is available.
- *Data collection instrument(s) in all languages they were administered.*
- *Abbreviations and other conventions.* Both variable names and variable labels will contain abbreviations. Ideally, these should be standardized.
- *Recode logic*. An audit trail of the steps involved in creating recoded variables.
- *Coding instruments.* Rules and definitions used for coding the data.
- *Related publications.* Citations to publications based on the data, by the principal investigators or others.

References

Ackermann-Piek, D., Silber, H., Daikeler, J., Martin, S., & Edwards, B. (2020). Interviewer Training Guidelines of Multinational Survey Programs: A Total Survey Error Perspective. *Methods, Data, Analyses*, *14*(1), 26.

Acquadro, C., Conway, K., Hareendran, A., Aaronson, N., Issues, E. R., & Group, Q. of L. A. (ERIQA). (2008). Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value in Health*, *11*(3), 509–521.

Acquadro, C., Patrick, D. L., Eremenco, S., Martin, M. L., Kuliś, D., Correia, H., Conway, K., & Research, I. S. for Q. of L. (2018). Emerging good practices for translatability assessment (TA) of patient-reported outcome (PRO) measures. *Journal of Patient-Reported Outcomes*, *2*(1), 8.

Afrobarometer Survey. (2014). *Round 6 Survey Manual*. http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Ahrendt, D., & MacGoris, S. (2018). *Improvements in Survey Quality over Time: Lessons Learnt from Eurofound's Pan-European Surveys Since 1995*. Comparative Survey Design and Implementation Workshop, Limerick, Ireland.

Alcser, K., Clemens, J., Holland, L., Guyer, H., & Hu, M. (2016). *Interviewer recruitment, selection, and training* [Guidelines for Best Practice in Cross-Cultural Surveys]. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Andreenkova, A. (2015). *Measuring acquiescence in different cultures: Results of experiments with translation and scale types*. Comparative Survey Design and Implementation, London.

Aquilino, W. S. (1993). Effects of spouse presence during the interview on survey responses concerning marriage. *Public Opinion Quarterly*, *57*(3), 358–376.

Aquilino, W. S., Wright, D. L., & Supple, A. J. (2000). Response Effects Due to Bystander Presence in CASI and Paper-and-Pencil Surveys of Drug Use and Alcohol Use. *Substance Use & Misuse*, *35*(6–8), 845–867. https://doi.org/10.3109/10826080009148424

Atkinson, P., Delamont, S., Cernat, A., Williams, R., & Sakshaug, J. W. (Eds.). (Forthcoming). *Sage Research Methods Foundations: An Encylclopeadia*. Sage.

Baker, R., Blumberg, S., Brick, J. M., Couper, M. P., Courtright, M., Dillman, D., Frankel, M. R., Garland, P., Groves, R., Kennedy, C., Krosnick, J. A., Lee, S., Lavrakas, P. J., Link, M., Piekarski, L., Rao, K., Rivers, D., Thomas, R. K., & Zahs, D. (2010). *AAPOR Report: Online Panels* [American Association for Public Opinion Research Report].

Baker, R., Brick, J. M., Bates, N., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). *Non-Probability Sampling* [American Association for Public Opinion Research Report].

Baldacci, E., Japec, L., & Stoop, I. (2016). Improving the comparability dimension of European Statistics by minimizing unnecessary variation. *European Conference on Quality in Official Statistics (Q 2016), Madrid (31 May–3 June)*.

Bargmeyer, B. E., and Gillman, D. W. "Metadata standards and metadata registries: An overview." *International Conference on Establishment Surveys II, Buffalo, New York*. 2000.

Bauer, J. J. (2016a). Biases in Random Surveys. In *Journal of Survey Statistics and Methodology* (pp. 263–287).

Bauer, J. (2016b). *Errors in Random Route Samples*. 1st SERISS Survey Experts Forum Workshop, Munich, Germany.

Beaumont, J.-F. (2005). On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment. *Survey Methodology*, *31*(2), 227–231.

Behr, D. (2009). *Translationswissenschaft und international vergleichende Umfrageforschung: Qualitätssicherung bei Fragebogenübersetzungen als Gegenstand einer Prozessanalyse. Bonn: GESIS. [Translation Research and Cross-National Survey Research: Quality Assurance in Questionnaire Translation from the Perspective of Translation Process Research]*.

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, *20*(6), 573–584.

Behr, D. (2018). Translating questionnaires for cross-national surveys: A description of a genre and its particularities based on the ISO 17100 categorization of translator competences. *Translation & Interpreting*, *10*(2), 5-20–20.

Behr, D. (Forthcoming). Computer-assisted migration research: What we can learn from the field of software localization for source questionnaire design and translation. In *Quantitative Migration Research in a Digitized World. Using Innovative Technology to Tackle Methodological Challenges*. IMISCOE-Springer.

Behr, D., & Braun, M. (2015). Satisfaction with the way democracy works: How respondents across countries understand the question. *Hopes and Anxieties: Six Waves of the European Social Survey*, 121–138.

Behr, D., & Shishido, K. (2016). The Translation of Measurement Instruments for Cross-Cultural Surveys. In C. Wolf, D. Joye, T. Smith, & Y. Fu, *The SAGE Handbook of Survey Methodology* (pp. 269–287). SAGE Publications Ltd.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). *Web probing—Implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions (Version 1.0)*.

Behr, D., & Scholz, E. (2011). Questionnaire translation in cross-national survey research: On the types and values of annotations. *Methoden, Daten, Analysen*, *5*(2), 157–179.

Behr, D., & Zabal, A. (2019). *A meeting report: OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments (2018)*. BioMed Central.

Benedict, R. (1946). Section of Anthropology: The Study of Cultural Patterns in European Nations. *Transactions of the New York Academy of Sciences*, *8*(8 Series II), 274–279.

Benford, F. (1938). The Law of Anomalous Numbers on JSTOR. *Proceedings of the American Philosophical Society*, *78*(4), 551–572.

Benítez, I., & Padilla, J.-L. (2014). Analysis of Nonequivalent Assessments Across Different Linguistic Groups Using a Mixed Methods Approach: Understanding the Causes of Differential Item Functioning by Cognitive Interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68.

Benstead, L. J. (2014). Does Interviewer Religious Dress Affect Survey Responses? Evidence from Morocco. *Politics and Religion*, 1–27.

Benstead, L. J., & Malouche, D. (2015). Interviewer Religiosity and Polling in Transitional Tunisia. *Midwest Political Science Association Annual Meeting, April*, 16–19.

Bergmann, M. & Schuller, K. (2019). Improving the efficiency of data quality back checks: A new procedure to prevent curbstoning. In Bergmann, M., Scherpenzeel, A., & Börsch-Supan, A. (eds.), *SHARE Wave 7 Methodology: Panel innovations and life histories*, MEA, Max Planck Institute for Social Law and Social Policy, Munich.

Bethlehem, J., Medrano, J. D., Groves, R. M., Gundelach, P., & Norris, P. (2008). Report of the Review Panel for the European Social Survey. *European Science Foundation, Standing Committee for Social Sciences (SCSC). Available from: Www. Europeansocialsurvey. Org. Accessed*, *22*, 2009.

Beullens, K., & Loosveldt, G. (2014). Interviewer effects on latent constructs in survey research. *Journal of Survey Statistics and Methodology*, *2*(4), 433–458.

Beullens, Koen, Matsuo, H., Loosveldt, G., & Vandenplas, C. (2014). Quality report for the European Social Survey, round 6. *London: European Social Survey ERIC*.

Beullens, K., & Loosveldt, G. (2016). Interviewer effects in the European Social Survey. *Survey Research Methods. Journal of the European Survey Research Association*, *10*(2), 103–118.

Beullens, K., Loosveldt, G., Denies, K., & Vandenplas, C. (2016). *Quality matrix for the European Social Survey, round 7*.

Beullens, K., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts? *Survey Methods: Insights from the Field (SMIF)*.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Biemer, P. P. (2016). Total survey error paradigm: Theory and Practice. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 122–141). SAGE.

Biemer, P. & Amaya, A. (2020). Total error frameworks for hybrid estimation and their applications. In C. Hill et al (eds*). Big data meets survey science. A collection of innovative methods*, Chapter 5, Wiley.

Biemer, P.P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley and Sons.

Biemer, P. P, Trewin, D., Bergdahl, H., & Japec, L. (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics*, *30*(3), 381–415. https://doi.org/10.2478/jos-2014-0022

Biemer, P. P., de Leeuw, E. D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., & West, B. T. (2017). *Total Survey Error in Practice*. John Wiley & Sons.

Billiet, J. (2016). What Does Measurement Mean in a Survey Context? In C. Wolf, D. Joye, T. Smith, & Y. Fu, *The SAGE Handbook of Survey Methodology* (pp. 193–209). SAGE Publications Ltd.

Bischke, B., Helber, P., Folz, J., Borth, D., & Dengel, A. (2019). Multi-task learning for segmentation of building footprints with deep neural networks. *2019 IEEE International Conference on Image Processing (ICIP)*, 1480–1484.

Blasius, J., & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research*, *52*, 479–493.

Blasius, J., & Thiessen, V. (2021). Perceived Corruption, Trust, and Interviewer Behavior in 26 European Countries. *Sociological Methods & Research* 50(2): 740-777.

Blom, A. G., Lynn, P., & Jäckle, A. (2008). *Understanding cross-national differences in unit non-response: The role of contact data*. ISER Working Paper Series.

Blom, A. G., De Leeuw, E. D., & Hox, J. (2011). Interviewer effects on nonresponse in the European Social Surveys. *Journal of Official Statistics*, 3–53.

Blom, A. G. (2016). Survey fieldwork. *The Sage Handbook of Survey Methodology*. Los Angeles: Sage, 382–396.

Boeije, H., & Willis, G. B. (2013). The cognitive interviewing reporting framework (CIRF): Towards the harmonization of cognitive testing reports. *European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(3), 558–565.

Bolaños-Medina, A., & González-Ruiz, V. (2012). Deconstructing the translation of psychological tests. *Meta: Journal Des Traducteurs/Meta: Translators' Journal*, *57*(3), 715–739.

Börsch-Supan, A., Brugiavini, A., Jürges, H., Kapteyn, A., Mackenbach, J., Siegrist, J., & Weber, G. (2008). First results from the Survey of Health, Ageing and Retirement in Europe (2004-2007). *Starting the Longitudinal Dimension. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA)*.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design—For market research, political polls, and social and health questionnaires*. Jossey-Bass.

Braun, M., & Mohler, P. P. (2003). Background variables. In *Cross-Cultural Survey Methods* (pp. 101–115). John Wiley and Sons.

Braun, M., Behr, D., Kaczmirek, L., & Bandilla, W. (2014). Evaluating cross-national item equivalence with probing questions in web surveys. In *Improving survey methods: Lessons from recent research* (pp. 184–200). Routledge.

Braun, M., Behr, D., & Díez Medrano, J. (2018). What do respondents mean when they report to be "citizens of the world"? Using probing questions to elucidate international differences in cosmopolitanism. *Quality & Quantity*, *52*(3), 1121–1135. https://doi.org/10.1007/s11135-017-0507-6

Braun, M., & Müller, W. (1997). Measurement of education in comparative research. *Comparative Social Research*, *16*, 163–201.

Bredl, S., Winker, P., & Kötschau, K. (2008). *A statistical approach to detect cheating interviewers*. Discussion Paper.

Brick, J. M. (2014). Explorations in non-probability sampling using the web. *Proceedings of the Conference on beyond Traditional Survey Taking: Adapting to a Changing World*, 1–6.

Brick, J. M., & Tourangeau, R. (2017). Responsive Survey Designs for Reducing Nonresponse Bias. *Journal of Official Statistics*, *33*(3), 735–752. https://doi.org/10.1515/jos-2017-0034

Brick, J. M., & Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59. https://doi.org/10.1177/0002716212456834

Brislin, R. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216.

Brislin, R. (1980). Translation and Content Analysis of Oral and Written Materials. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of Cross-Cultural Psychology* (Vol. 2, pp. 389–444). Allyn & Bacon.

Buchanan, W., & Cantril, H. (1953). *How nations see each other: A study in public opinion*. Greenwood Press.

Bush, S. S., & Prather, L. (n.d.). *How Electronic Devices in Face-to-Face Interviews Change Survey Behavior: Evidence from a Developing Country*. Retrieved from http://www.laurenprather.org/uploads/2/5/2/3/25239175/bush_prather_electronic_devices_in_survey_interviews.pdf, 2020.

Bushery, J. M., Reichert, J. W., Albright, K. A., & Rossiter, J. C. (1999). *Getting More Bang from the Reinterview Buck: Identifying "At Risk" Interviewers* (Proceedings from Section on Survey Research Methods, pp. 316--320). American Statistical Association.

Buskirk, T., Bear, T., & Bareham, J. (2018). *Machine made sampling designs: Applying machine learning methods for generating stratified sampling designs*. Big Data Meets Survey Science Conference, Barcelona, Spain.

Butler, D. (2013). When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature*, *494*(7436), 155–157.

Cajka, J., Amer, S., Ridenhour, J., & Allpress, J. (2018). Geo-sampling in developing nations. *International Journal of Social Research Methodology*, *21*(6), 729–746.

Callegaro M., & Yang Y. (2017). The Role of Surveys in the Era of "Big Data". In Vannette D., Krosnick J. (eds) *The Palgrave Handbook of Survey Research*. Palgrave Macmillan, Cham.

Calvo, E. (2018). From translation briefs to quality standards: Functionalist theories in today's translation processes. *Translation & Interpreting*, *10*(1), 18-32–32. https://doi.org/10.12807/t&i.v10i1.639

Campanelli, P., Sturgis, P., & Purdon, S. (1997). *Can you hear me knocking? An investigation into the impact of interviewers on survey response rates*.

Carey, S. (2000). *Measuring adult literacy: The International Adult Literacy Survey (IALS) in the European context*. Office for National Statistics.

Caspar, R., Peytcheva, E., Yan, T., Lee, S., Liu, M., & Hu, M. (2016). Pretesting. *Cross-Cultural Survey Guidelines. August*.

Casterline, J., & Chidambaram, V. C. (1984). The presence of others during the interview and the reporting of contraceptive knowledge and use. *Survey Analysis for the Guidance of Family Planning Programs. Liege, Belgium: Ordina Editions*, 267–298.

Center for Capacity Building in Survey Methods and Statistics. (2018). Short Course Series in International and Cross-Cultural Surveys. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved Month, March 22, 2020, from https://ccb.isr.umich.edu.

Chen, J., Valliant, R. L., & Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), 657–681.

Chew, R. F., Amer, S., Jones, K., Unangst, J., Cajka, J., Allpress, J., & Bruhn, M. (2018). Residential scene classification for gridded population sampling in developing countries using deep convolutional neural networks on satellite imagery. *International Journal of Health Geographics*, *17*(1), 12.

Chidlow, A., Plakoyiannaki, E., & Welch, C. (2014). Translation in cross-language international business research: Beyond equivalence. *Journal of International Business Studies*, *45*(5), 562–582.

Cibelli Hibben, K. L., de Jong, J., Hu, M., Durow, J., & Guyer, H. (2016). *Study design and organizational structure* [Guidelines for Best Practice in Cross-Cultural Surveys]. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Ciepłuch, B., Jacob, R., Mooney, P., & Winstanley, A. C. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, 337.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons.

Colina, S., Marrone, N., Ingram, M., & Sánchez, D. (2017). Translation quality assessment in health research: A functionalist alternative to back-translation. *Evaluation & the Health Professions*, *40*(3), 267–293.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Sage Publications.

Couper, Mick P., & Lyberg, L. (2005). The use of paradata in survey research. *Proceedings of the 55th Session of the International Statistical Institute*.

Couper, M.P. (1998). Measuring survey quality in a CASIC environment. *American Statistical Association: Survey Research Methods Section*. http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf

Daas, P. J., & Puts, M. J. (2014). *Social media sentiment and consumer confidence*. ECB Statistics Paper.

Dale, A., Arber, S., & Procter, M. (1988). *Doing secondary analysis*. Unwin Hyman.

Dalenius, T. (1967). Nonsampling errors in census and sample surveys. Report no.5 in the research project Errors in Surveys. Stockholm University.

de Jong, J. A. (2019). Ethical Considerations in the Total Survey Error Context. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 665–682). John Wiley & Sons, Inc.

de Jong, J. A., Dorer, B., Lee, S., Yan, T., & Villar, A. (2019). Overview of questionnaire design and testing. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 115–137). John Wiley & Sons, Inc.

de Jong, J. A., Young-Demarco, Moaddel, M., & Gelfand, M. (2016). Best practices: Lessons from a Middle East survey research program. In *Values, Political Action, and Change in the Middle East and the Arab Spring* (pp. 295–323). Oxford University Press.

de Jong, J. A., & Cibelli Hibben, K. (2018). *European Quality of Life Survey 2016: Quality Assessment*. Eurofound.

de Jong, J. A., Mneimneh, Z. N., & Moaddel, M. (2017). *Measuring Third-Party Presence During Face-to-Face Interviews: Respondent and Interviewer Predictors & Effect on Reporting Sensitive Attitudes in Jordan and Turkey*. International Workshop on Comparative Survey Design and Implementation, Mannheim, Germany.

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.). (2008). *International Handbook of Survey Methodology*. Taylor & Francis Group/Lawrence Erlbaum Associates.

de Leeuw, E. D., Suzer-Gurtekin, Z. T., & Hox, J. J. (2019). The Design and Implementation of Mixed-mode Surveys. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, 387–409.

Dean, E., Caspar, R., McAvinchey, G., Reed, L., & Quiroz, R. (2007). Developing a low-cost technique for parallel cross-cultural instrument development: The question appraisal system (QAS-04). *International Journal of Social Research Methodology*, *10*(3), 227–241.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and Cultural Appropriateness of Survey Material in Large-Scale Assessments. *Implementation of Large-Scale Education Assessments*, 168–192.

Dorer, B. (Forthcoming). *Advance Translation as a Means of Improving Source Questionnaire Translatability? Findings from a Think-Aloud Study for French and German*. Frank & Timme.

Douglas, S. P., & Craig, C. S. (2007). Collaborative and iterative translation: An alternative approach to back translation. *Journal of International Marketing*, *15*(1), 30–43.

Eckman, Stephanie, & Koch, A. (2019). Interviewer involvement in sample selection shapes the relationship between response rates and data quality. *Public Opinion Quarterly*, *83*(2), 313–337.

Ehling, M. (2003). Harmonising data in official statistics. In *Advances in cross-national comparison* (pp. 17–31). Springer.

Ehling, M., & Rendtel, U., et al. (2006). *Synopsis. Research Results of Chintex-Summary and Conclusions*.

Epstein, J., Osborne, R. H., Elsworth, G. R., Beaton, D. E., & Guillemin, F. (2015). Cross-cultural adaptation of the Health Education Impact Questionnaire: Experimental study showed expert committee, not back-translation, added value. *Journal of Clinical Epidemiology*, *68*(4), 360–369. https://doi.org/10.1016/j.jclinepi.2013.07.013

European Social Survey. (2018a). *ESS9—2018 Documentation Report: Appendix A6 (Classifications and Coding Standards)*. Norwegian Centre for Research Data.

European Social Survey (2018b). *ESS Round 9 Translation Guidelines*. London: ESS ERIC Headquarters.

European Social Survey. (2020). *ESS9—2018 Documentation Report, Edition 1.3*. Norwegian Centre for Research Data.

European Union. (2018). *Standard Eurobarometer 90—Autumn 2018: Public Opinion in the European Union, First Results*. https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/84930

Eurostat. (2009). *ESS Standards for Quality Reports*. Eurostat. https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-ESQR_FINAL.pdf

Eurostat. (2017). *Quality report of the European Union Labour Force Survey, 2015*. https://ec.europa.eu/eurostat/documents/3859598/10276257/KS-GQ-19-012-EN-N.pdf/f7c1b8dd-7246-01a3-dcec-328d2f38acd9

Eurostat. (2019). *Statistical requirements compendium | 2019 edition*. https://ec.europa.eu/eurostat/documents/3859598/10276257/KS-GQ-19-012-EN-N.pdf/f7c1b8dd-7246-01a3-dcec-328d2f38acd9

Fetvadjiev, V. H., Meiring, D., Van de Vijver, F. J., Nel, J. A., & Hill, C. (2015). The South African Personality Inventory (SAPI): A culture-informed instrument for the country's main ethnocultural groups. *Psychological Assessment*, *27*(3), 827.

Finn, A., & Ranchhod, V. (2013). *Genuine Fakes: The prevalence and implications of fieldworker fraud in a large South African survey*. Southern Africa Labour and Development Research Unit.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fitzgerald, R. (2015). *Striving for quality, comparability and transparency in cross-national social survey measurement: Illustrations from the European Social Survey (ESS)* (Unpublished Doctoral Thesis, City University London). http://openaccess.city.ac.uk/14487/

Fitzgerald, R., Winstone, L., & Prestage, Y. (2014). *A Versatile tool? Applying the Cross-national Error Source Typology (CNEST) to triangulated pre-test data*. FORS Working Paper Series, paper 2014-2.

Fitzgerald, R., & Zavala-Rojas, D. (2020). A Model for Cross-National Questionnaire Design and Pretesting. In P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 493–520).

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation* (Vol. 38). Sage Publications.

Frankovic, K., Johnson, T., & Stavrakantonaki, M. (2017). *Freedom to Conduct Opinion Polls*. ESOMAR/WAPOR. https://wapor.org/wp-content/uploads/ESOMA-WAPOR_Freedom-to-Conduct-Opinion-Polls-Final-incl-edits.pdf

Gallup Europe. (2010). *Quality Assessment of the 5th European Working Conditions Survey*. Eurofound.

Gambier, Y. (2016). Translations| Rapid and Radical Changes in Translation and Translation Studies. *International Journal of Communication*, *10*, 20.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, *69*(1), 124–157.

Geisen, E., & Romano Bergstrom, J. (2017). Usability and Usability Testing. *Usability Testing for Survey Research*, 1–19.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Girres, J.-F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, *14*(4), 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

Goerman, P. L. (2017). *Cognitive Interview Standards and Guidelines at the U.S. Census Bureau: Implementation and use across Languages.* Washington Statistical Society Seminar on Implementing the New OMB Cognitive Interviewing Standards and Guidelines.

Goerman, P., Meyers, M., & García Trejo, Y. (2018). The Place of Expert Review in Translation and Questionnaire Evaluation for Hard-to-Count Populations in National Surveys. *GESIS Symposium on" Surveying the Migrant Population: Consideration of Linguistic and Cultural Aspects"*, *19*, 29–41.

Göpferich, S., & Jääskeläinen, R. (2009). Process research into the development of translation competence: Where are we, and where do we need to go? *Across Languages and Cultures*, *10*(2), 169–191.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID)A Web Facility that Tests Question Comprehensibility. *Public Opinion Quarterly*, *70*(1), 3–22. https://doi.org/10.1093/poq/nfj012

Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing survey data. In *Survey methods in multinational, multicultural and multiregional contexts* (pp. 315–332). John Wiley & Sons.

Granda, P., & Blasczyk, E. (2016). *Data Harmonization* [Guidelines for Best Practice in Cross-Cultural Surveys]. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Groves, R. M. (2004). *Survey Errors and Survey Costs* (Vol. 536). Wiley-Interscience.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861–871.

Groves, R. M. (2018). In Defense of Disciplines. *The Provost's Blog*. https://blog.provost.georgetown.edu/author/bgroves/page/10/

Groves, R. M. (2019). Personal communication with Lars Lyberg.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. John Wiley & Sons.

Groves, R. M., Dillman, D. A., Etinge, J. L., & Little, R. J. A. (2002). *Survey nonresponse*. John Wiley & Sons.

Groves, R. M., & Harris-Kojetin, B. A. (2017). Federal Statistics, Multiple Data Sources, and Privacy Protection. Washington, D.C.: The National Academies Press.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 439–457.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons Inc.

Groves, R.M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

Gryna, F. M., & Juran, J. M. (2001). *Quality planning and analysis: From product development through use*. McGraw-Hill New York.

Gutmann, M. P., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal*, *3*, 209–221.

Hagell, P., Hedin, P.-J., Meads, D. M., Nyberg, L., & McKenna, S. P. (2010). Effects of Method of Translation of Patient-Reported Health Outcome Questionnaires: A Randomized Study of the Translation of the Rheumatoid Arthritis Quality of Life (RAQoL) Instrument for Sweden. *Value in Health*, *13*(4), 424–430. https://doi.org/10.1111/j.1524-4733.2009.00677.x

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets—Mordechai Haklay, 2010. *Environment and Planning B: Planning and Design*, *37*(4), 682–703.

Hansen, S. E., Benson, G., Bowers, A., Pennell, B.-E., Lin, Y., Duffey, B., Hu, M., & Cibelli Hibben, K. L. (2016). *Survey quality* [Guidelines for Best Practice in Cross-Cultural Surveys].

Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Hansen, M., Hurwitz, W. & Pritzker, L. (1964). The estimation andinterpretation of gross differences and simple response variance. In C.R. Rao (ed) Contributions to statistics, 111-136, Oxford: Pergamon Press.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–54). Wiley-Interscience.

Harkness, J. A. (2008). Comparative survey research: Goals and challenges. In E. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 56–77). Psychology Press Taylor & Francis Group.

Harkness, J. A., & Schoua-Glusberg, A. S. (1998). Questionnaires in translation. In *Cross-cultural survey equivalence* (p. 2007). ZUMA.

Harkness, J. A., Van de Vijver, F. J. R., & Johnson, T. P. (2003). Questionnaire design in comparative research. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-Cultural Survey Methods* (pp. 19–34). John Wiley & Sons.

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). John Wiley & Sons, Inc.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010a). Designing questionnaires for multipopulation research. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 31–57). John Wiley & Sons.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., & Lyberg, L. E. (2010b). *Survey methods in multicultural, multinational, and multiregional contexts*. John Wiley & Sons.

Harkness, J. A., Stange, M., Cibelli, K. L., Mohler, P. Ph., & Pennell, B.-E. (2014). Surveying cultural and linguistic minorities. In R. Tourangeau, B. Edwards, T. P. Johnson, K. Wolter, & N. Bates (Eds.), *Hard-to-survey populations*. Cambridge University Press.

Harkness, J. A., Bilgen, I., Cazar, A. C., Huang, L., Miller, D., Stange, M., & Villar, A. (2016). Questionnaire Design. In *Guidelines for Best Practice in Cross-Cultural Surveys*. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Harter, R., Eckman, S., English, N., & O'Muircheartaigh, C. (2010). Applied sampling for large-scale multistate area probability designs. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 169–195). Emerald Group Publishing.

Heeb, J.-L., & Gmel, G. (2001). Interviewers' and respondents' effects on self-reported alcohol consumption in a Swiss health survey. *Journal of Studies on Alcohol*, *62*(4), 434–442.

Heeringa, S. G. (2017). *Survey-assisted modeling: Integration of sample survey designs and methods with big data systems*. 5th School on Sampling and Survey Methodology, Cuiaba, Brazil. https://www.ufmt.br/dest/arquivos/c7aa221a2bd387d99f3495a1d78af6ec.pdf

Heeringa, S. G., & O'Muircheartaigh, C. (2010). Sample design for cross-cultural and cross-national survey programs. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 251–267). John Wiley & Sons.

Hoelter, L., Pienta, A., & Lyle, J. (2016). *Data Preservation, Secondary Analysis, and Replication: Learning from Existing Data*.

Hoffmeyer-Zlotnik, J. H. (2016). *Standardisation and Harmonisation of Socio-Demographic Variables (Version 2.0)*.

Hoffmeyer-Zlotnik, J. H., & Wolf, C. (2003). *Advances in cross-national comparison: A European working book for demographic and socio-economic variables*. Kluwer Academic/Plenum Publishers.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. SAGE Publications.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79–125.

Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2008). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. *Advances in Telephone Survey Methodology*, *1*, 499–528.

Hood, C. C., & Bushery, J. M. (1997). *Getting More Bang from the Reinterview Buck: Identifying "At Risk" Interviewers* (Proceedings from Section on Survey Research Methods, pp. 820–824). American Statistical Association.

Hox, J. J., & De Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity*, *28*(4), 329–344.

Hubbard, F., Lin, Y., Zahs, D., & Hu, M. (2016). *Sample Design* [Guidelines for Best Practice in Cross-Cultural Surveys]. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Hyder, S., Bilal, L., Akkad, L., Lin, Y., Al-Habeeb, A., Al-Subaie, A., Shahab, M., Binmuammar, A., Al-Tuwaijr, F., Kattan, N., & Altwaijri, Y. (2017). Evidence-based guideline implementation of quality assurance and quality control procedures in the Saudi National Mental Health Survey. *International Journal of Mental Health Systems*, *11*(1), 60. https://doi.org/10.1186/s13033-017-0164-0

Inciardi, J. A. (1981). *The drugs-crime connection*. Sage Publications Beverly Hills.

International Monetary Fund (IMF). (2012). *Data quality assessment framework*. https://dsbb.imf.org/content/pdfs/dqrs_Genframework.pdf

International Organization for Standardization. (2015a). *ISO 17100:2015 Translation Services*. ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/91/59149.html

International Organization for Standardization. (2015b). *ISO 9001:2015 Quality management systems—Requirements*. https://www.iso.org/standard/62085.html

International Organization for Standardization ISO. (2019). *ISO 20252:2019 Market, opinion and social research, including insights and data analytics—Vocabulary and service requirements*. https://www.iso.org/standard/73671.html

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition)*. https://www.intestcom.org/

Jääskeläinen, R. (2010). Are all professionals experts. *Translation and Cognition*, *15*, 213–227.

Jäckle, A., Lynn, P., Sinibaldi, J., & Tipping, S. (2013). The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods*, *7*, 1–15.

Japec, L. (2005). *Quality issues in interview surveys: Some contributions* [PhD Thesis]. Statistiska institutionen.

Japec, L., Kreuter, F., Berg, M., Biemer, P. P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). *AAPOR Report: Big Data* [American Association for Public Opinion Research Report].

Japec, L. & Lyberg, L. (forthcoming). Big data initiatives in official statistics. In Hill, C., Biemer, P., Buskirk, T., Japec, L., Kirchner, A., Kolenikov, S., & Lyberg, L. (Eds). Big Data Meets Survey Science. Wiley.

Ji, L.-J., Zhang, Z., & Nisbett, R. E. (2004). Is It Culture or Is It Language? Examination of Language Effects in Cross-Cultural Research on Categorization. *Journal of Personality and Social Psychology, 87*(1), 57–65.

Johnson, T. P. (1998). Approaches to establishing equivalence in cross-cultural and cross-national survey research. In *ZUMA-Nachrichten Spezial, 3*: 1–40.

Johnson, T. P. (2019). *Overall Goals of 3MC Research*. Annual Comparative Survey Design and Implementation Workshop, Warsaw, Poland.

Johnson, T. P., O'Rourke, D., & Chavez, N. (1997). Social cognition and response to survey questions among culturally diverse populations. In *Survey measurement and process quality* (pp. 87–113). Wiley & Sons.

Johnson, T. P., & Parsons, J. A. (1994). Interviewer effects on self-reported substance use among homeless persons. *Addictive Behaviors*, *19*(1), 83–93.

Johnson, T. P., & Braun, M. (2016). *Challenges of comparative survey research* (SAGE Handbook of Survey Methodology). Sage.

Johnson, T. P., Pennell, B.-E., Stoop, I. A., & Dorer, B. (2019a). The Promise and Challenge of 3MC Research. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, 3–12.

Johnson, T. P., Pennell, B.-E., Stoop, I. A., & Dorer, B. (Eds.). (2019b). *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)*. John Wiley & Sons, Inc.

Jowell, R. (1998). How comparative is comparative research? *American Behavioural Scientist, 42*, 168–177.

Kallas, J., & Linardis, A. (2010). A documentation model for comparative research based on harmonization strategies. *IASSIST Quarterly*, *32*(1), 12–12.

Kalton, G., Lyberg, L., & Rempp, J.-M. (1998). *Review of methodology. In Adult literacy in OECD Countries: Technical report on the first International Adult Literacy Survey*. U.S. National Center for Education Statistics.

Kelley, J., Krishna, K. S., & Lai, J. (2015). *Designing a new data capture method: Usability study of the Instagram app as a data collection tool*. American Association for Public Opinion Research (AAPOR) conference, Hollywood, FL.

Kenett, R.S. & Shmueli, G. (2014). Journal of the Royal Statistical Society, A, 177 (1), 3-27.

Kessler, R. C., & Üstün, T. B. (2008). *The WHO world mental health surveys: Global perspectives on the epidemiology of mental disorders*. Cambridge University Press; Published in collaboration with the World Health Organization.

Kirgis, N. G., & Lepkowski, J. M. (2013). Design and management strategies for paradata-driven responsive design: Illustrations from the 2006–2010 National Survey of Family Growth. In *Improving surveys with paradata* (pp. 121–144). John Wiley & Sons, Inc.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.

Kleiner, B., Pan, Y., & Bouic, J. (2009). The impact of instructions on survey translation: An experimental study. *Survey Research Methods*, *3*, 113–122.

Koch, A. (2019). With-in household selection of respondents. In T.P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology* (pp. 93–109). John Wiley & Sons, Inc.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods, 1*, 55–67.

Kolsrud, K., Rød, L.-M., & Segadal, K. U. (2019). Linking auxiliary data to survey data: Ethical and legal challenges in Europe and the United States. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 683–704). John Wiley & Sons, Inc.

Kołczyńska, M. (2014). Representation of Southeast European countries in international survey projects: Assessing data quality. *ASK. Research & Methods*, *23*, 57–78.

Kołczyńska, M. (2018). "Sampling Schemes and Survey Quality in Cross-national Surveys" Presentation delivered at the Comparative Survey Design and Implementation (CSDI) International Workshop, March 26-28 Limerick, Ireland.

Kołczyńska, M., & Schoene, M. (2019). Survey Data Harmonization and the Quality of Data Documentation in Cross-national Surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods* (pp. 963–984). John Wiley & Sons, Ltd.

Koller, M., Kantzer, V., Mear, I., Zarzar, K., Martin, M., Greimel, E., Bottomley, A., Arnott, M., Kuliś, D., & TCA-SIG, T. I. (2012). The process of reconciliation: Evaluation of guidelines for

translating quality-of-life questionnaires. *Expert Review of Pharmacoeconomics & Outcomes Research, 12*(2), 189–197. https://doi.org/10.1586/erp.11.102

Kresja, E. A., Davis, M. C., & Hill, J. M. (1999). Evaluation of the quality assurance falsification interview used in the Census 2000 dress rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 365–640.

Kreuter, F. (Ed.). (2013). *Improving Surveys with Paradata: Analytical Uses of Process Information*. John Wiley & Sons, Inc.

Kreuter, F. (2017). Getting the Most Out of Paradata. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 193–198). Springer.

Kreuter, F., & Casas-Cordero, C. (2010). Paradata. In German Data Forum (RatSWD) (Ed.), *Building on Progress* (1st ed., pp. 509–530). Verlag Barbara Budrich; JSTOR. https://doi.org/10.2307/j.ctvbkk43d.31

Kreuter, F., Couper, M. P., & Lyberg, L. E. (2010). The use of paradata to monitor and manage survey data collection. In *Proceedings of the Joint Statistical Meetings, American Statistical Association* (pp. 282–296).

Kreuter, F., & Olson, K. (2013). Paradata for nonresponse error investigation. *Improving Surveys with Paradata: Analytic Uses of Process Information, 2*, 13–42.

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS, 32*(3), 283–291.

Landrock, U. (2017). How Interviewer Effects Differ in Real and Falsified Survey Data: Using Multilevel Analysis to Identify Interviewer Falsifications. *Methods, Data, Analyses*, *11*(2), 26.

Lavrakas, P. J. (1992). *Chicagoans' attitudes towards and experience with select sexual issues: Harassment, discrimination, AIDS, homosexuality*. Northwestern University Survey Laboratory.

Le, K. T., Brick, J. M., Diop, A., & Alemadi, D. (2013). Within-household sampling conditioning on household size. *International Journal of Public Opinion Research*, *25*(1), 108–118.

Lee, T., & Pérez, E. O. (2014). The Persistent Connection Between Language-of-Interview and Latino Political Opinion. *Political Behavior*, *36*(2), 401–425. https://doi.org/10.1007/s11109-013-9229-1

Lee, S., Keusch, F., Schwarz, N., Liu, M., & Suzer-Gurtekin, T. Z. (2019). Cross-cultural comparability of response patterns of subjective probability questions. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 457–475). John Wiley & Sons, Inc.

Lepkowski, J. (2005). Non-observation error in household surveys in developing countries. *Household Sample Surveys in Developing and Transition Countries: United Nations*, 149–70.

Lepkowski, J. M., Mosher, W. D., Groves, R. M., West, B. T., Wagner, J., & Gu, H. (2013). *Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth*.

Li, J., Brick, J. M., Tran, B., & Singer, P. (2011). Using Statistical Models for Sample Design of a Reinterview Program. *Journal of Official Statistics*, *372*(3), 433–450.

Lipps, O. (2007). Interviewer and respondent survey quality effects in a CATI panel. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *95*(1), 5–25.

Liu, M., Suzer-Gurtekin, T. Z., Keusch, F., & Lee, S. (2019). Response styles in cross-cultural surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 477–499). John Wiley & Sons, Inc.

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the European Social Survey. *Journal of Official Statistics*, *33*(2), 409–426.

Lyberg, L. E. (2012). Survey quality. *Survey Methodology*, 107–130.

Lyberg, L. E., & Biemer, P. P. (2008). Quality assurance and quality control in surveys. In *International handbook of survey methodology*. Lawrence Erlbaum Associates.

Lyberg, L. E., Hanover, L., Cibelli Hibben, K. L., & Pennell, B.-E. (2018). Applying Total Survey Error and survey process quality to the Programme for International Assessment of Adult Competencies. *Quality Assurance in Education*, *26*(2), 153–168.

Lyberg, L. E., & Stukel, D. M. (2010). Quality assurance and quality control in cross-national comparative studies. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 225–249). John Wiley & Sons Inc.

Lyberg, L. & Stukel, D. (2017). The roots and evolution of the total survey error concept. In P. Biemer et al (eds) Total survey error in practice, Chapter 1, 3-2, Wiley.

Lyberg, L. E., Japec, L., & Tangur, C. (2019). Prevailing issues and the future of comparative surveys. In T.P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 1055–1077). John Wiley & Sons, Inc.

Lyberg, L. E., & Weisberg, H. (2016). Total survey error: A paradigm for survey methodology. . In C. Wolf, D. Joye, T. Smith, & Y. Fu, *The SAGE Handbook of Survey Methodology* (pp. 27-42). SAGE Publications Ltd.

Lynn, P. & Lugtig, P. (2017). Total survey error for longitudinal surveys. In P.Biemer, et al (eds) *Total survey error in practice*, Chapter 13, 279-298, Wiley.

Lynn, P., Japec, L., & Lyberg, L. E. (2006). What's so special about cross-national surveys? *International Workshop on Comparative Survey Design and Implementation (CSDI)*, *12*, 7–20.

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M.-J. (2018). The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension. *Public Opinion Quarterly*, *82*(4), 707–744. https://doi.org/10.1093/poq/nfy038

Maineri, A., Scherpenzeel, A., Bristle, J., Pflüger, S.-M., Mindarova, I., Butt, S., Zins, S., Emery, T., & Luijkx, R. (2017). *Report on the use of sampling frames in European studies*.

Maitland, A., & Presser, S. (2016). How accurately do different evaluation methods predict the reliability of survey questions? *Journal of Survey Statistics and Methodology*, *4*(3), 362–381.

Marsden, P. (2011). Survey methods for network data. In J. Scott & P. Carrington (eds) *The SAGE Handbook of social network analysis*, 370-388. London: Sage Publications.

Malter, F. (2017). *"Curbstoning": Case study of an elaborate interviewer falsification scheme and new procedures to prevent interviewer fabrication*. European Survey Research Association, Lisbon, Portugal.

McKenna, S. P., & Doward, L. C. (2005). The Translation and Cultural Adaptation of Patient-Reported Outcome Measures. *Value in Health*, *8*(2), 89–91.

Mehrbrodt, T., Gruber, S., & Wagner, M. (2017). Scales and multi-item indicators. *Munich, Germany: Survey of Health, Ageing and Retirement in Europe*.

Meitinger, K. (2017). Necessary but InsufficientWhy Measurement Invariance Tests Need Online Probing as a Complementary Tool. *Public Opinion Quarterly*, *81*(2), 447–472.

Menold, N. (2014). The influence of sampling method and interviewers on sample realization in the European Social Survey. *Survey Methodology*, *40*(1), 105–123.

Miller, J., Słomczynski, K. M., & Schoenberg, R. J. (1981). Assessing comparability of measurement in cross-national research: Authoritarian-conservatism in different sociocultural settings. *Social Psychology Quarterly*, 178–191.

Miller, K. (2019). Conducting Cognitive Interviewing Studies to Examine Survey Question Comparability. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, 203.

Miller, K., Fitzgerald, R., Caspar, D. M., Gray, M., Nunes, C., & Padilla, J. L. (2008). *Design and analysis of cognitive interviews for cross-national testing*. International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts, Berlin, Germany.

Mneimneh, Z.N. (2012). Interview privacy and social conformity effects on socially desirable reporting behavior: Importance of cultural, individual, question, design and implementation factors. *Doctoral Dissertation, The University of Michigan*.

Mneimneh, Z., Tourangeau, R., Pennell, B.-E., Heeringa, S. G., & Elliott, M. R. (2015). Cultural variations in the effect of interview privacy and the need for social conformity on reporting sensitive information. *Journal of Official Statistics*, *31*(4), 673–697.

Mneimneh, Z., de Jong, J., Cibelli Hibben, K., & Moaddel, M. (2018). Do I Look and Sound Religious? Interviewer Religious Appearance and Attitude Effects on Respondents' Answers. *Journal of Survey Statistics and Methodology*.

Mneimneh, Z., Lyberg, L. E., Sharma, S., Vyas, M., Sathe, D. B., Malter, F., & Altwaijri, Y. (2019). Case studies on monitoring interviewer behavior in international and multinational surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 731–770). John Wiley & Sons, Inc.

Mneimneh, Z., de Jong, J., & Altwaijri, Y. (2020). Why Do Interviewers Vary on Interview Privacy and Does Privacy Matter? In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective*. CRC Press.

Mneimneh, Z., Cibelli Hibben, K., de Jong, J., & Kelley, J. (Forthcoming). Comparative Surveys. In P. Atkinson, S. Delamont, A. Cernat, R. Williams, & J. W. Sakshaug (Eds.), *Sage Research Methods Foundations: An Encyclopedia*. Sage.

Mohler, P. Ph. (2006). *Sampling from a universe of items and the de-machiavellization of questionnaire design* (Beyond the Horizon of Measurement - Festschrift in Honor of Ingwer Borg (ZUMA-Nachrichten Spezial,10)). ZUMA. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial10.pdf

Mohler, P. Ph, & Uher, R. (2003). *Documenting comparative surveys for secondary analysis*. New York: Wiley.

Mohler, P. Ph., Pennell, B.-E., & Hubbard, F. (2008). *Survey documentation: Toward professional knowledge management in sample surveys* (International Handbook of Survey Methodology). Lawrence Erlbaum Associates.

Mohler, P.Ph., & Johnson, T. P. (2010). Equivalence, Comparability, and Methodological Progress. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, 17–29.

Mohler, P. Ph, Hansen, S. E., Pennell, B.-E., Thomas, W., Wackerow, J., & Hubbard, F. (2010). A survey process quality perspective on documentation. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, 299–314.

Mohler, P. Ph., Dorer, B., de Jong, J., & Hu, M. (2016). *Translation: Overview* [Guidelines for Best Practice in Cross-Cultural Surveys]. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Montalvo, J. D., Seligson, M. A., & Zechmeister, E. J. (2018). Data Collection in Cross-national and International Surveys: Latin America and the Caribbean. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, 569–582.

Morganstein, D., & Marker, D. A. (1997). Continuous quality improvement in statistical agencies. In *Survey measurement and process quality* (pp. 475–500). John Wiley & Sons.

Moskowitz, J. M. (2004). Assessment of Cigarette Smoking and Smoking Susceptibility among YouthTelephone Computer-Assisted Self-Interviews versus Computer-Assisted Telephone Interviews. *Public Opinion Quarterly*, *68*(4), 565–587. https://doi.org/10.1093/poq/nfh040

Murphy, J., Baxter, R. K., Eyerman, J., & Cunningham, D. (2004). *A system for detecting interviewer falsification*. American Association for Public Opinion Research, Phoenix, AZ.

Murray, T. S., Kirsch, I. S., & Jenkins, L. (1998). *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey*. US Department of Education, Office of Educational Research and Improvement ….

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Nida, E. A. (1964). *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill Archive.

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation* (Vol. 8). Brill.

Niu, J., & Hedstrom, M. (2008). Documentation evaluation model for social science data. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 11–11.

Noelle, E. (1963). *Umfragen in der Massengesellschaft: Einführung in die Methoden der Demoskopie*. Verlag für Demoskopie.

Nord, C. (2014). *Function and Loyalty in Bible Translation*. Apropos of Ideology.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010, May 16). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Fourth International AAAI Conference on Weblogs and Social Media*. Fourth International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1536

Oldendick, R. W., Bishop, G. F., Sorenson, S. B., & Tuchfarber, A. J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics*, *4*(4), 307–318.

Oleksiyenko, O., Wysmulek, I., & Vangeli, A. (2019). Identification of Processing Errors in Cross-national Surveys. In Johnson, T.P, Pennell, B.-E., Stoop, I.A., & Dorer, B. (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 985–1010). John Wiley & Sons, Inc.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., McCarthy, J., O'Brien, E., Opsomer, J., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., & Wagner, J. (2019). *AAPOR Report: Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys* [American Association for Public Opinion Research Report].

O'Muircheartaigh, C., & Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *161*(1), 63–77.

Orlowski, R. A., Antoun, C., Carlson, R., & Hu, M. (2016). *Tenders, Bids, and Contracts* [Guidelines for Best Practice in Cross-Cultural Surveys]. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Orten, H., Norland, S., & Butt, S. (2018). *The Questionnaire Design and Documentation Tool (QDDT): A DDI based tool for assisting questionnaire design teams in their work*. https://zenodo.org/record/2530046#.XHWo1c9KiRu

Padilla, J.-L., Benítez, I., & Van de Vijver, F. (2019). Addressing Equivalence and Bias in Cross-cultural Survey Research Within a Mixed Methods Framework. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 45–64). John Wiley & Sons, Inc.

Pan, Y., & de la Puente, M. (2005). *Census Bureau guidelines for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed* (Research Report Series #2005-06). US Bureau of the Census.

Pan, Y., Landreth, A., Park, H., Hinsdale-Shouse, M., & Schoua-Glusberg, A. (2010). Cognitive interviewing in non-English languages: A cross-cultural perspective. In *Survey methods in multinational, multiregional, and multicultural contexts*. John Wiley & Sons.

Pan, Y., Sha, M., & Park, H. (2019). *The Sociolinguistics of Survey Translation*. Routledge.

Park, H., & Goerman, P. L. (2019). Setting up the cognitive interview task for non-English speaking participants in the U.S. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 227–249). John Wiley & Sons, Inc.

Park, H., Sha, M. M., & Willis, G. (2016). Influence of English-language Proficiency on the Cognitive Processing of Survey Questions. *Field Methods*, *28*(4), 415–430.

Pennell, B.-E., Harkness, J. A., Levenstein, R., & Quaglia, M. (2010). Challenges in cross-national data collection. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 269–298). John Wiley & Sons.

Pennell, B.-E., Deshmukh, Y., Kelley, J., Maher, P., & Tomlin, D. (2014). Disaster research: Surveying displaced populations. In R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, & N. Bates (Eds.), *Hard-to-survey populations*. Cambridge University Press. http://www.cambridge.org/us/academic/subjects/psychology/psychology-research-methods-and-statistics/hard-survey-populations

Pennell, B.-E., & Cibelli Hibben, K. L. (2016). Surveying in multicultural and multinational contexts. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *Sage Handbook of survey methodology* (pp. 157–177). Sage Publications, Inc.

Pennell, B.-E., Cibelli Hibben, K. L., Lyberg, L., Mohler, P. Ph., & Worku, G. (2017). A Total Survey Error perspective on surveys in multinational, multiregional, and multicultural contexts. In P. Biemer, E. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, C. Tucker, & B. T. West (Eds.), *Total survey error in practice*. John Wiley & Sons Inc.

Petrakos, M., Kleideri, M., & Ieromnimon, A. (2010). *Quality Assessment of the 2nd European Quality of Life Survey*. Eurofound.

Peyrard, N., Sabbadin, R., Spring, D., Brook, B., & Mac Nally, R. (2013). Model-based adaptive spatial sampling for occurrence map construction. *Statistics and Computing*, *23*(1), 29–42.

Peytchev, A. (2013). Consequences of Survey Nonresponse. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 88–111. https://doi.org/10.1177/0002716212461748

Peytcheva, E. (2019). Can the language of survey administration influence respondents' answers. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 325–340). John Wiley & Sons, Inc.

Peytcheva, E. (2020). The Effect of Language of Survey Administration on the Response Formation Processes. In M. Sha & T. Gabel (Eds.), *The essential role of language in survey research*. RTI Press.

Peytcheva, E., & Groves, R. M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, *25*(2), 193.

PIAAC. (2014). *PIAAC Technical Standards and Guidelines*. OECD. https://www.oecd.org/skills/piaac/PIAAC-NPM(2014_06)PIAAC_Technical_Standards_and_Guidelines.pdf

Pickery, J., & Loosveldt, G. (2000). Modeling interviewer effects in panel surveys: - An application. *Survey methodology*, *26*(2), 189–198.

Pickery, J., & Loosveldt, G. (2002). A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse. *Quality & Quantity*, *36*(4), 427–437.

Platek, R., & Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, *17*(1), 1–20.

Porras, J., & English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. *Proceedings of the American Statistical Association (Survey Research Methods Section)*, 4223–4228.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. Ò., Martin, E. A., & Martin, J. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley & Sons.

Przeworski, A., & Teune, H. (1966). Equivalence in cross-national research. *Public Opinion Quarterly*, *30*(4), 551–568.

Rammstedt, B., Beierlein, C., Brähler, E., Eid, M., Hartig, J., Kersting, M., Liebig, S., Lukas, J., Mayer, A.-K., Menold, N., Schupp, J., & Weichselgartner, E. (2015). *Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research* (Working Paper No. 245). RatSWD Working Paper. https://www.econstor.eu/handle/10419/107203

Revilla, M., Ochoa, C., & Toninelli, D. (2016). PCs versus Smartphones in answering web surveys: Does the device make a difference? *Survey Practice*, *9*(4).

Revilla, M., Zavala-Rojas, D., & Saris, W. E. (2016). Creating a good question: How to use cumulative experience. *The SAGE-Handbook of Survey Methodology*, 236–254.

Riley, J. (2017). *Understanding metadata: What is metadata, and what is it for?* National Information Standards Organization. http://www.niso.org/publications/understanding-metadata-riley

Rizzo, L., Brick, J. M., & Park, I. (2004). A minimally intrusive method for sampling persons in random digit dial surveys. *The Public Opinion Quarterly*, *68*(2), 267–274.

Robbins, M. (2019). New frontiers in detecting data fabrication. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 771–805). John Wiley & Sons, Inc.

Rokkan, S. (1969). Cross-national survey research: Historical, analytical, and substantive contexts. In S. Rokkan, S. Verba, J. Viet, & E. Almsay (Eds.), *Comparative Survey Analysis* (pp. 5–55). Mouton.

Roller, M. R., & Lavrakas, P. J. (2015). *Applied Qualitative Research Design: A Total Quality Framework Approach*. Guilford Publications.

Ruggles, S. (2018). The Importance of Data Curation. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 303–308). Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_39

Salmon, C. T., & Nichols, J. S. (1983). The Next-Birthday Method of Respondent Selection. *Public Opinion Quarterly*, *47*(2), 270–276. https://doi.org/10.1086/268785

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.

Saris, W., Oberski, D., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I., & Gruner, T. (2011). *The development of the program SQP 2.0 for the prediction of the quality of survey questions* (RECSM Working Paper 24).

Schäfer, C., Schräpler, J. P., Müller, K. R., & Wagner, G. G. (2004). *Automatic identification of faked and fraudulent interviews in surveys by two different methods*. European Conference on Quality and Methodology in Official Statistics, Mainz, Germany. http://www.diw.de/documents/publikationen/73/diw_01.c.42515.de/dp441.pdf

Scherpenzeel, A. C., Maineri, A., Bristle, J., Pflüger, S.-M., Mindarova, I., Butt, S., Zins, S., Emery, T., & Luijkx, R. (2017). *Report on the use of sampling frames in European studies* (Deliverable 2.1 of the SERISS Project).

Schneider, S. L. (2007). Measuring educational attainment in cross-national surveys: The case of the European Social Survey. *EDUC Workshop of the EQUALSOC Network, Dijon*, 22–24.

Schneider, S. L., Briceno-Rosas, R., Ortmanns, V., & Herzing, J. M. (2018). Measuring migrants' educational attainment: The CAMCES tool in the IAB-SOEP migration sample. *Surveying the Migrant Population: Consideration of Linguistic and Cultural Issues*, 43–74.

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 389–410.

Schnepf, S. (2018). *Insights into survey errors of large scale educational achievement surveys*. JRC Working Papers in Economics and Finance.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*(1), 180–211.

Scholz, E., & Heller, M. (2009). *ISSP Study monitoring 2007* (GESIS Technical Reports, 2009/5). GESIS. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/TechnicalReport_09-5.pdf

Schoua-Glusberg, A. S. (1992). *Report on the translation of the questionnaire for the national treatment improvement evaluation study*. National Opinion Research Center.

Schoua-Glusberg, A. S. (2004). *Decisions Translators Make: A Case for Detailed Specifications*. Second International Workshop on Comparative Survey Design and Implementation, Paris, France.

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In J. A. Harkness, M. Braun, B. Edwards, T. P.

Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 175–190). John Wiley & Sons, Inc.

SDR Team. (2019). Survey Data Recycling as an Analytic Framework for Survey Data Reprocessing. Part 1: Introduction and Theoretical Overview. Presentation prepared by I. Tomescu-Dubrow & K. M. SŁomczyński, asc.ohio-state.edu/dataharmonization/about/events/building-multi-source-databases-december-2019/.

Seligson, M. A., & Moreno Morales, D. E. (2018). Improving the Quality of Survey Data Using CAPI Systems in Developing Countries. In L. R. Atkeson & R. M. Alvarez (Eds.), *The Oxford Handbook of Polling and Survey Methods* (pp. 207–219). Oxford University Press.

SERISS. (2020). *Synergies for Europe's Research Infrastructures in the Social Sciences*. https://seriss.eu/

Sha, M., Hsieh, Y. P., & Goerman, P. L. (2018). Translation and visual cues: Towards creating a road map for limited English speakers to access translated Internet surveys in the United States. *Translation & Interpreting*, *10*(2), 142–158.

Sha, M., Park, H., Pan, Y., & Kim, J. (2020). Cross-Cultural Comparison of Focus Groups as a Research Method. In M. Sha & T. Gabel (Eds.), *The essential role of language in survey research*. (pp.151-179). RTI Press.

Shreve, G. M. (2002). Knowing translation: Cognitive and experiential aspects of translation expertise from the perspective of expertise studies. *Translation Studies: Perspectives on an Emerging Discipline*, 150–171.

Silber, H., Stark, T. H., Blom, A. G., & Krosnick, J. A. (2019). Implementing a multinational study of questionnaire design. In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 161–179). Wiley & Sons.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. In *Survey Nonresponse* (pp. 163–178). John Wiley & Sons.

Sinibaldi, J., Durrant, G. B., & Kreuter, F. (2013). Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly, 77*(S1), 173–193.

SŁomczyński, K. M., Tomescu-Dubrow, I., Jenkins, J. C., Kołczyńska, M., Powałko, P., Wysmułek, I., Oleksiyenko, O., Zieliński, M. W., & Dubrow, J. K. (2016). *Democratic Values and Protest Behavior: Harmonization of Data from International Survey Projects*. IFiS Publishers.

Słomczyński, K. M., Jenkins, J. C., Tomescu-Dubrow, I., Kołczyńska, M., Wysmułek, I., Oleksiyenko, O., Powałko, P., & Zieliński, M. W. (2017a). *SDR 1.0 Master Box* [Data set]. https://doi.org/10.7910/DVN/VWGF5Q

Słomczynski, K. M., Powalko, P., & Krauze, T. (2017b). Non-unique Records in International Survey Projects: The Need for Extending Data Quality Control. *Survey Research Methods*, *11*(1), 1–16. https://doi.org/10.18148/srm/2017.v11i1.6557

Słomczyński, K. M., & Tomescu-Dubrow, I. (2019). Basic Principles of Survey Data Recycling. In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey*

Methods: Multinational, Multiregional, and Multicultural Contexts (3MC) (pp. 937–962). Wiley & Sons.

Smith, S. N., Fisher, S. D., & Heath, A. (2011). Opportunities and challenges in the expansion of cross-national survey research. *International Journal of Social Research Methodology*, *14*(6), 485–502.

Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-Cultural Survey Methods* (pp. 69–92). John Wiley & Sons.

Smith, T. W. (2004). Developing and Evaluating Cross-National Survey Instruments. In Stanley Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (1st ed., pp. 431–452). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471654728

Smith, T. W. (2007). Survey non-response procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods*.

Smith, T. W. (2008). Making translation an integrated, scientific component of cross-national survey research. *American Sociological Association Conference, Boston*.

Smith, T. W. (2010). The globalization of survey search. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P.-Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 477–484). Wiley Hoboken, NJ.

Smith, T. W. (2011). Refining the total survey error perspective. *International Journal of Public Opinion Research*, 464–484.

Smith, T. W. (2015). A Review of Survey Data-Collection Modes. *Sociological Theory and Methods*.

Smith, T. W. (2019a). Improving Multinational, Multiregional, and Multicultural (3MC) Comparability Using the Total Survey Error (TSE) Paradigm. In T.P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 13-43). John Wiley & Sons, Inc.

Smith, T. W. (2019b). Optimizing questionnaire design in cross-national and cross-cultural surveys. In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 473–492). John Wiley & Sons, Inc.

Smith, T. W. (2019c). A Preliminary Report on the Validation and Verification of Interviews and Interviewers. Unpublished International Social Survey Programme report.

Smyth, J. D. (2016). Designing Questions and Questionnaire. *The Sage Handbook of Survey Methodology*, 218.

Son, J. (2018). Back translation as a documentation tool. *Translation & Interpreting*, *10*(2), 89–100.

Sorensen, N., & Oyserman, D. (2009). Collectivism, effects on relationships. In *Encyclopedia of Human Relationships*. SAGE Publications. http://www .sageereference.com.proxy.lib.umich.edu/humanrelationships/Article_n80.htm

Stoop, I., Matsuo, H., Koch, A., & Billiet, J. (2010). Paradata in the European Social Survey: Studying nonresponse and adjusting for bias. *Proceedings of the Survey Research Methods Section, ASA*, 407–421.

Stoop, I., & Koch, A. (2013). Data collection as a scientific process: Process control and process quality in the European Social Survey. *Understanding Research Infrastructures in the Social Sciences. Zürich: Seismo*, 145–57.

Stoop, I., Koch, A., Halbherr, V., Loosveldt, G., & Fitzgerald, R. (2016). *Field Procedures in the European Social Survey Round 8: Guidelines for Enhancing Response Rates and Minimising Nonresponse Bias*. ESS ERIC Headquarters. http://www.europeansocialsurvey.org/docs/round8/methods/ESS8_guidelines_enhancing_response_rates_minimising_nonresponse_bias.pdf

Stoop, Ineke, Briceno-Rosas, R., Koch, A., & Vandenplas, C. (2018). *Data Falsification in the European Social Survey*. European Social Survey.

Sudman, S. (1966). New approaches to control of interviewing costs. *Journal of Marketing Research*, 55–61.

Sundgren, B. (1973). *An infological approach to data bases.* [PhD Thesis]. Stockholm University.

Sundgren, B. (1995). *Guidelines for the Modelling of Statistical Data and Metadata* [Publisehd as Guidelines from the United Nations Statistical Division, New York]. sites.google.com/site/bosundgren/my-life

Survey Research Center. (2016). *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Tarnia, J., Rosa, E., & Scott, L. P. (1987). *An Empirical Comparison of the Kish and the Most Recent Birthday Method for Selecting a Random Household Respondent in Telephone Surveys*. Annual meetings of the American Association for Public Opinion Research, Hershey, PA.

Tessler, M. A. (2011). *Public opinion in the Middle East: Survey research and the political orientations of ordinary citizens*. Indiana University Press.

Tofangsazi, B., & Lavryk, D. (2018). We Coded the Documentation of 1700+ Surveys…. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, *42*(2), 27–31.

Tomescu-Dubrow, I., & Słomczyński, K. M. (2016). Harmonization of cross-national survey projects on political behavior: Developing the analytic framework of survey data recycling. *International Journal of Sociology*, *46*(1), 58–72.

Tomescu-Dubrow, I., Słomczynski, K. M., & Kołczyńska, M. (2017). Quality Controls and Their Application to Substantive Analyses of Data from International Survey Projects. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, *3*(1), 9–13.

Tomescu-Dubrow, I., & Granda, P. (2019). *A New Look at Types of Probes for Testing Translated Instruments*. Survey Documentation in 3MC Surveys, Warsaw, Poland.

Tortora, R. D., Srinivasan, R., & Esipova, N. (2010). The Gallup World Poll. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 535–543).

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.

Tourangeau, Roger, Edwards, B., Johnson, T. P., Bates, N., & Wolter, K. M. (Eds.). (2014). *Hard-to-Survey Populations*. Cambridge University Press.

Turner, C. F., Rogers, S. M., Miller, H. G., Miller, W. C., Gribble, J. N., Chromy, J. R., Leone, P. A., Cooley, P. C., Quinn, T. C., & Zenilman, J. M. (2002). Untreated Gonococcal and Chlamydial Infection in a Probability Sample of Adults. *JAMA*, *287*(6), 726–733.

United Nations. (2005). *Household surveys in developing and transition countries*. United Nations, Department of Economic and Social Affairs. https://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf

Uskul, Ayse K., & Oyserman, D. (2006). Question Comprehension and Response: Implications of Individualism and Collectivism. In Y.-R. Chen (Ed.), *National Culture and Groups* (Vol. 9, pp. 173–201). Emerald Group Publishing Limited.

Uskul, A.K., Oyserman, D., & Schwarz, N. (2010). Cultural emphasis on honor, modesty, or self-enhancement: Implications for the survey-response process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 191–201). John Wiley & Sons, Inc.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045.

Vannieuwenhuyze, J. T. A., Loosveldt, G., & Molenberghs, G. (2014). Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models. *Journal of Official Statistics*, *30*(1), 1–21.

Vannieuwenhuyze, J. T., & Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods & Research*, *42*(1), 82–104.

van den Brakel, J. A., Vis-Visschers, R., & Schmeets, J. J. G. (2006). An Experiment with Data Collection Modes and Incentives in the Dutch Family and Fertility Survey for Young Moroccans and Turks. *Field Methods*, *18*(3), 321–334. https://doi.org/10.1177/1525822X06287533

Vardigan, M., Granda, P., & Hoelter, L. (2016). Documenting surveys across the data life cycle. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 443-569). Sage.

Vassallo, R., Durrant, G. B., Smith, P. W. F., & Goldstein, H. (2015). Interviewer effects on non-response propensity in longitudinal surveys: A multilevel modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 83–99.

Vehovar, V., Slavec, A., & Berzelak, N. (2012). Costs and Errors in Fixed and Mobile Phone Surveys. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 277–295). Springer New York. https://doi.org/10.1007/978-1-4614-3876-2_16

Verba, S., & Almond, G. A. (1963). *The civic culture: Political attitudes and democracy in five nations*. Princeton University Press.

Verba, Sidney, Nie, N. H., & Kim, J. (1978). *Participation and Political Equality: A Seven-Nation Comparison*. CUP Archive.

Vila, J., Cervera, J., & Carausu, F. (2013). *Quality assessment of the third European Quality of Life Survey*. Eurofound.

Villar, A., & Fitzgerald, R. (2017). Using mixed modes in survey research: Evidence from six experiments in the ESS. *Europe: Values and Identities*, 273–310.

Wagner, J., & Stoop, I. A. L. (2019). Comparing Nonresponse and Nonresponse Biases in Multinational, Multiregional, and Multicultural Contexts. In T.P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 807–833). John Wiley & Sons, Inc.

Wagner, M., Philip, J. T., & Jürges, H. (2019). Questionnaire innovations in the seventh wave of SHARE. *SHARE Wave 7 Methodology: Panel Innovations and Life Histories*. http://www.share-project.org/data-documentation/methodology-volumes.html

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001

West, B. T. (2011). Paradata in survey research. *Survey Practice*, *4*(4), 1–8.

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 211–225.

West, B. T., & Kreuter, F. (2013). Factors Affecting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth. *Public Opinion Quarterly*, *77*(2), 522–548.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94–104.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.

Willis, G, B., & Lessler, J. T. (1999). *Question appraisal system BRFSS-QAS: A guide for systematically evaluating survey question wording*. Research Triangle Institute: CDC/NCCDPHP/Division of Adult and Community Health Behavioral Surveillance Branch.

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods*, *23*, 331–341.

Wolf, C., Joye, D., Smith, T. W., & Fu, Y. (2016a). *The SAGE Handbook of Survey Methodology*. SAGE.

Wolf, C., Schneider, S. L., Behr, D., & Joye, D. (2016b). Harmonizing survey questions between cultures and over time. *The SAGE Handbook of Survey Methodology*, 502–524.

Wuyts, C., & Loosveldt, G. (2019). *Quality matrix for the European Social Survey, Round 8*. European Social Survey.

Wyer Jr, R. S. (2013). Culture and information processing: A conceptual integration. In *Understanding Culture* (pp. 431–455). Psychology Press.

Yan, T., & Hu, M. (2019). Examining Translation and Respondents' Use of Response Scales in 3MC Surveys. In *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 501–518). Wiley & Sons.

Yang, Y., Harkness, J. A., Chin, T., & Villar, A. (2010). Response styles and culture. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 203–223). John Wiley & Sons.

Yang, H. L., Lunga, D., & Yuan, J. (2017). Toward country scale building detection with convolutional neural network using aerial images. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 870–873.

Zavala-Rojas, D. (2018). Exploring Language Effects in Crosscultural Survey Research: Does the Language of Administration Affect Answers About Politics? *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (Mda), 12*(1), 127–150.

Zavala-Rojas, D., Saris, W. E., & Gallhofer, I. (2019). Preventing differences in translated survey items using the Survey Quality Predictor. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 357–384). John Wiley & Sons, Inc.

Zieliński, M. W., Powałko, P., & Kołczyńska, M. (2019). The Past, Present, and Future of Statistical Weights in International Survey Projects: Implications for Survey Data Harmonization. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, 1035–1052.